# DATA JOURNALISM ASSESSMENT MALAYSIA

## Data Journalism in Constrained Environment

LICENSING
Disseminated under Creative Commons Attribution 4.0 International License (CC-BY-4.0) (http://creativecommons.org/licenses/by/4.0/)

# Acknowledgments

Sinar Project and Internews wish to acknowledge Dr. Yusminar Yunus and the Global Data Barometer team for recent research and additional up to date sectoral data availability and sources for Malaysia. Malaysian journalists featured in the case studies Aidila Razak, Yao-Hua Law and Ida Lim for their data journalism stories and sharing their insights and experiences. Kuek Ser Kuang Keng from DataN and the participants of Internews Data Journalism and Investigative Journalism Workshops whose challenges in trying to find data for their stories helped guide the direction of this assessment. James McKinney for Popolo, CoST – the Infrastructure Transparency Initiative and Open Contracting Partnership for the development of data standards and guides that help simplify data gathering and reporting for persons of interest, infrastructure projects and procurement.

Finally, we would like to thank the individuals and organisations whose work and previous research contributed towards the understanding of data journalism and media landscape.

**Author**
Khairil Yusof

**Collaborating Partner –** Sinar Project
**Coordinators –** Nadiah Rosli, Nurulnabillah Ahmad Hijazu (Internews)
**Reviewer -** Eva Constantaras (Internews)

**Graphic Designer –** Victor Andin
**Proofreader and BM Translator –** Farah Draim

**For more information:**
www.sinarproject.org
www.internews.org

# Table of Contents


**Summary** 5

**Key Points/Findings** 5

**Summary of Findings** 5

**State of Open Data and Data Journalism in Malaysia** 6

Limited availability of open data 6

No Right to Information 7

Constraints on Media Freedom and Freedom of Expression 9

Data Availability and Sources 9

General 9

Demographics 10

Company Information 10

Crime 10

Transportation 10

Elections and Legislative Data 11

Land 11

Political Integrity 11

Public Finance 11

Climate Action and Environment 12

Health & COVID-19 13

International Trade 13

**Case Studies** 14

Case Study 1 - Kini News Lab COVID-19 Tracker 14

Journalist/Organisation 14

Description 14

Challenges 14

Methodology and Impact 15

Sources of Data / Data Generated 15

Case Study 2 - Child Marriages in Malaysia 16

Journalist/Organisation 16

Description 16

Challenges 16

Methodology and Techniques 17

Sources of Data / Data Generated 18

Case Study 3 - Evaluating Forest Management in Peninsular Malaysia 19
Journalist/Organisation 19
Description 19
Challenges 20
Methodology and Techniques 20
Sources of Data / Data Generated 20
Case Study 4 - Politically Exposed Persons 21
Journalist/Organisation 21
Description 21
Challenges 21
Methodology and Techniques 22
Sources of Data / Data Generated 23

**Best Practices in Data Journalism in Constrained Environments 24**
Organize Data Structure Using Open Data Standards and Hypothesis 24
Example: Water Disruption 24
Open Data Standards - Electoral Data 25
Open Data Standards - Infrastructure and Procurement 27
Searching Government Documents and Websites 30
Government Documents, 30
Government Reports 30
Securities Commissions 30
Investigative Journalism Resources 31

**Biodata of Author 32**

**End Note 33**

# Summary

## Key Points/Findings

Lack of quality granular open data.

Restrictive legal environment for right to information and freedom of expression make getting access to data difficult and puts journalists and sources at risk of persecution

Malaysian journalists are applying innovative methods in sourcing and generating their own datasets for data journalism despite these challenges

## Summary of Findings

There is limited open data across all sectors available for Malaysian journalists, making it a difficult and challenging environment for data journalism. What open data is available, often lacks granularity and is often incomplete. Given a story idea, it is unlikely that a large quality dataset can be found, ready for use for a data driven story. Data, however is still available, but fragmented across multiple sources local and international, and as unstructured data in documents and graphical infographics.

Journalists face legal restrictions in sourcing for data, with no national Right to Information law, overly broad Official Secrets Act and severe criminal sanctions with up to a year's jail for public officials sharing information, even when such information is not filed as a state secret. When publishing a story, journalists face additional challenges of overly broad laws restricting freedom of expression for both online and print media.

Despite these restrictions, Malaysian journalists have developed innovative methods in finding and sourcing their data needed for award winning data journalism from COVID-19 to deforestation. This includes, creating their own data structures in spreadsheets, and populating it from a variety of sources, and at times, becoming the source of public interest data through these efforts long after the data driven story has been published.

This difficult environment blurs the boundaries between investigative and data journalism. In addition to providing readily available sources of data, this assessment report also shares case study examples of data journalism on health, environment, human rights and anti-corruption by leading local journalists, and a compilation of techniques and tips in finding and sourcing data needed for data journalism in Malaysia.

# State of Open Data and Data Journalism in Malaysia

*Open without Freedom*

Poor Data Availability

No Right to Information

Constraints on Media Freedom and Freedom of Expression

## Limited availability of open data

Open Data Barometer 2016

The Malaysian government has had an open data initiative since 2015, which includes a central government open data portal data.gov.my, a government wide circular providing guidance on the need for open data[1] as well as standard license by government agencies for publication of open data. Superseding directive in 2021, it provided even more detailed guidance on publication of open data, including open data requests between government agencies.[2] As of 2021, the portal publishes over 12 thousand datasets from 394 agencies and departments.[3] Government websites also have standard disclaimer and page on open data.

Open data surveys from Open Data Barometer (2016)[4] and Global Open Data Index (2015)[5], find that availability of quality, granular open data for key government sectors is very poor. Sectors such as environment and health, areas which Malaysian journalists attending data journalism workshops wanted to do their stories on were ranked poorly in terms of availability.

Civil society groups and researchers continue to note that granular, complete, timely and open data, the same criteria needed for data journalism, is often not available or free. Data published on government websites were also often not machine readable and published in the form of PDFs. [6]

**2. Jenayah hartabenda yang dilaporkan kepada polis mengikut jenis, Malaysia, 1999 – 2004**

| Jenis Jenayah | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 |
|---|---|---|---|---|---|---|
| **Jumlah** | **147,958** | **145,569** | **136,079** | **128,199** | **133,525** | **134,596** |
| **Pecah rumah dan curi** | 35,936 | 32,913 | 28,452 | 25,265 | 25,789 | 24,904 |
| **Siang hari** | 9,401 | 8,675 | 7,449 | 6,821 | 6,928 | 6,550 |
| **Malam hari** | 26,535 | 24,238 | 21,003 | 18,444 | 18,861 | 18,354 |
| **Kecurian kenderaan** | 53,069 | 57,775 | 60,049 | 74,891 | 64,300 | 65,076 |
| **Lori/van** | 3,485 | 3,698 | 4,306 | 4,570 | 5,551 | 4,892 |
| **Kereta** | 6,196 | 7,278 | 8,520 | 8,544 | 8,537 | 8,624 |
| **Motosikal/skuter** | 41,905 | 45,903 | 47,223 | 47,137 | 50,212 | 51,560 |
| **Basikal** | 1,483 | 896 | - | 14,640 | - | - |
| **Ragut** | - | - | 14,368 | - | 15,798 | 11,536 |
| **Lain-lain kecurian** | 58,953 | 54,881 | 33,210 | 28,043 | 27,638 | 33,080 |

Caption: Non-machine readable data as PDFs. Criminal property reported to the police according to types, Malaysia, 1999 – 2004. Source: Royal Malaysia Police

Punca : Polis DiRaja Malaysia

These problems are compounded for data journalists, because granular data needs to be combined across sectors such as education levels against local demographic data to provide more meaningful insight for stories.

Limited availability of useful open data for Malaysian data journalists, requires that additional techniques in finding and generating data needed for stories, will need to be applied.

# No Right to Information

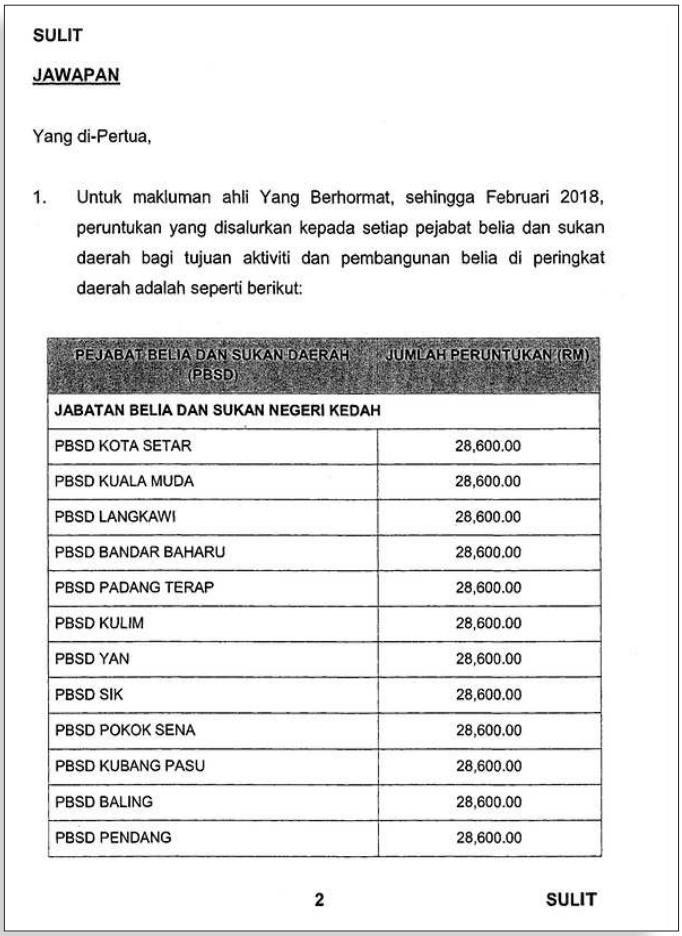
Data availability in Malaysia, also faces considerable challenges due to a lack of a national Right to Information (RTI) law, compounded with multiple laws and regulations with broad restrictions on publication of government information with heavy punitive penalties. This creates a closed by default environment for access to data and information, counter to the goals of open data.

The Official Secrets Act 1972 provides wide discretionary powers for the government to restrict availability of data. Broad terms of any official document along with conferring powers to any government official, provides for broad discretionary power to define any government document including data as an official secret, and carries with it a penalty of a year's imprisonment.[7] This has lead it to being abused in practice, with even public published documents such as parliamentary reports or local council papers in the State of Selangor[8]. This also creates a dangerous environment for journalists, because even linking a classified document can lead to arrest.[9]

The tabling of a state level Freedom of Information law for the state of Selangor, is one example of how an RTI law can have a positive impact on making data and information open by default. Since the tabling of the enactment, local council papers and other state government documents are no longer classified as an official secret.

*"(1) Whoever discloses any information or matter which has been obtained by him in the performance of his duties or the exercise of his functions under any written law shall be punished with fine of not more than one million Ringgit, or with imprisonment for a term which may extent to one year, or with both.* - Penal Code Section 203A

Additionally another law, Penal Code Section, 203A, further puts public officers at risk of a RM1 million fine or a year's imprisonment or both, for disclosure of any information or matter obtained while on the job.[10] This law provides an additional impediment to journalists trying to obtain information or data on request, as it puts their government sources at risk. As was the case for journalist Boo Su Lyn for Code Blue, who was questioned by police for investigations for citing a public government report in reporting a hospital fire.[11]



SULIT

**JAWAPAN**

Yang di-Pertua,

1. Untuk makluman ahli Yang Berhormat, sehingga Februari 2018, peruntukan yang disalurkan kepada setiap pejabat belia dan sukan daerah bagi tujuan aktiviti dan pembangunan belia di peringkat daerah adalah seperti berikut:

| PEJABAT BELIA DAN SUKAN DAERAH (PBSD) | JUMLAH PERUNTUKAN (RM) |
|---|---|
| **JABATAN BELIA DAN SUKAN NEGERI KEDAH** | |
| PBSD KOTA SETAR | 28,600.00 |
| PBSD KUALA MUDA | 28,600.00 |
| PBSD LANGKAWI | 28,600.00 |
| PBSD BANDAR BAHARU | 28,600.00 |
| PBSD PADANG TERAP | 28,600.00 |
| PBSD KULIM | 28,600.00 |
| PBSD YAN | 28,600.00 |
| PBSD SIK | 28,600.00 |
| PBSD POKOK SENA | 28,600.00 |
| PBSD KUBANG PASU | 28,600.00 |
| PBSD BALING | 28,600.00 |
| PBSD PENDANG | 28,600.00 |

2 SULIT

Caption: Public parliamentary reply marked as secret.

# Restrictive licences/ paywalls for data and public information

-literary work includes-

...

***but does not include*** *official texts of the Government or statutory bodies of a legislative or regulatory nature, or judicial decisions, or political speeches and political debates, or speeches delivered in the*

*course of legal proceedings, and the official translation thereof*

Malaysian Copyright Act 1987

Malaysian Copyright Act 1987 (Act 332)[12] states that official texts of the Government or statutory bodies are in the public domain and not copyrightable. Despite this provision, civil society groups and researchers have raised concerns that a lot of data is not free.[13] Many agencies are also required by regulation to charge a fee for data provision when the data coverage is larger than a specific scope in terms of aggregation, volume or time series.[14] This has led to government documents that legally should be in the public domain, assigned a copyright and only accessible via a paywall such as the case with State Gazettes.

## Constraints on Media Freedom and Freedom of Expression

At the publication stage, journalists in Malaysia also faced a variety of challenges both legal and editorial, with press freedom categorised as a difficult situation and dropping down 18 places in the World Press Freedom Report 2021.[15] Broad laws that restrict the work of journalists include the Sedition Act and the Communications and Multimedia Act. Editorial independence is also limited with print and television media owned by the business arms of political parties, leading to self-censorship by journalists.[16]

Recent cases include expelling two of Al-Jazeera's Australian journalists, and seven journalists questioned on 4th August, 2021 for reporting of treatment and conditions of migrant workers. A migrant worker who appeared in the story, was arrested and deported.[17]

The World Bank Open Data Readiness Assessment report also found in interviews that the Malaysian government did not regularly identify media/journalists as a potential user of open data.[18] The report also recommended that online media, due to their diversity in comparison to traditional media, is likely to have the best starting point for using data.

## Data Availability and Sources

This section covers data availability and sources by sector. Note that due to lack of open data availability in general for Malaysia, sources are likely to be inconsistent in terms of coverage, granularity and quality.

As data availability and sources are constantly being updated, a live updated version of the data sources section is available online. (https://docs.google.com/spreadsheets/d/1mg0jZiAWfTM4se8kowL9ECQa5TM-eIxWHW2L7xR-pzA/edit#gid=0)

Due to limited data, especially granular, journalists should not limit sourcing their data from specific sector sources only. Granular data on demographics of indigenous groups for example, can be found in environmental assessment reports, instead of general demographic statistics.

A similar approach can also be taken in trying to source Malaysian data from international data sources such as the World Bank.

## General

**MAMPU Open Data Government Website**
https://www.data.gov.my/

**Department of Statistics Malaysia Open Data**
https://www.dosm.gov.my/v1/index.php?r=column3/accordion&menu_id=aHhRYUpWS3B4VXlYaVBOeUF0WFpWUT09

**MysIDC**
Key datasets by sector
https://mysidc.statistics.gov.my/index.php?lang=en#

**eStatistik**
https://newss.statistics.gov.my

Over 3,400 additional datasets in machine readable and PDF format can also be found for download at https://newss.statistics.gov.my which requires a free online account registration process. A search feature for datasets that are freely downloadable is available via the "Free Download" section.

Caption: Datasets and data reports for eStatistik, can be searched under the Free Download section.

# Demographics

## Census data

| Data | Source | Notes |
|---|---|---|
| Households | Census Data Open Data Portal | Poor granularity |
| Education | Pendidikan - Clusters - MAMPU | Good |
| Employment | Census Data Open Data Portal | Good |
| Birthplace & Residence | Census Data Open Data Portal | Good |
| Population Distribution | Census Data Open Data Portal | Good |

## Company Information

| Data | Source | Notes |
|---|---|---|
| Beneficial Ownership | MY-Data SSM | Payment required. Bulk data download planned, but currently not available. |
| Company Registra | MY-Data SSM | Payment required. Bulk data download planned, but currently not available. |

## Crime

| Data | Source | Notes |
|---|---|---|
| Crime Statistics | Parliamentary Documents<br>Department of Statistics Malaysia<br><br>https://www.data.gov.my/data/en_US/group/jenayah | Not granular. Limited years and some data are in PDF format. |

## Transportation

| Data | Source | Notes |
|---|---|---|
| Bus Timetable | Pengangkutan - Clusters - MAMPU | Not comprehensive, and does not cover all major service providers.<br><br>No routes. |
| Rail | Pengangkutan - Clusters - MAMPU | Not comprehensive, and does not cover all major service providers.<br><br>No routes. |

## Elections and Legislative Data

| Data | Source | Notes |
|---|---|---|
| Electoral Results | https://www.data.gov.my/data/ms_MY/organization/election-commission-of-malaysia-spr | Not granular and inconsistent. Only for recent elections. |
| Map Boundaries | https://github.com/TindakMalaysia | Not up to date. |

## Land

| Data | Source | Notes |
|---|---|---|
| Land Tenure | Pemilikan Tanah - Clusters - MAMPU | Inconsistent data availability. |
| Existing Land Use | Pemilikan Tanah - Clusters - MAMPU<br><br>World Bank<br>https://data.worldbank.org/indicator/AG.LND.AGRI.ZS?locations=MY<br><br>i-Plan Geo Portal<br>https://iplan.townplan.gov.my/public/geoportal?view=zoning | Inconsistent data availability. |

## Political Integrity

| Data | Source | Notes |
|---|---|---|
| Political Finance | N/A | Not available. |
| Asset Declarations | https://mydeclaration.sprm.gov.my/ | Limited and aggregated monthly income only. |
| Lobbying Register | N/A | Not available. |
| Public Consultation Data | N/A | Not available. |
| RTI Performance | N/A | Not available. |

## Public Finance

| Data | Source | Notes |
|---|---|---|
| Budget | Bajet - Clusters - MAMPU | Inconsistent and limited data availability. |
| Expenditure | Perbelanjaan Kerajaan - Clusters - MAMPU | Inconsistent and limited data availability. |

## Climate Action and Environment

| Data | Source | Notes |
|---|---|---|
| Emission | MysDIC<br>https://mysidc.statistics.gov.my/index.php?lang=en#<br>Environment Cluster of Open Data Portal | National level granular data covering several years on MysDIC. |
| Biodiversity | MEWA Data | Only list of projects. |
| Vulnerability | Perhilitan Data<br>CITES Permit Data<br>Alam Sekitar - Clusters - MAMPU | Inconsistent and limited data availability. |
| Forest | MysDIC https://mysidc.statistics.gov.my/index.php?lang=en#<br>Annual Reports | National level by state data covering several years on MysDIC on forest coverage.<br>In PDF annual reports: https://www.forestry.gov.my/my/pusat-sumber/penerbitan/laporan-tahunan |
| Forest Fires | NASA Firms https://firms.modaps.eosdis.nasa.gov/active_fire/ | |

### TIPS: UN COMTRADE Data

Many environmental issues will involve trade of goods, services and commodities. Issues such as plastics, logging or fisheries products will involve import and export of such goods directly or indirectly.

If data for complete product is not available, try find out what goods and services are used in it's production.

First look up the HS Code definition for the goods in question and then search for the detailed open data which can be found on UN Comtrade website https://comtrade.un.org/data

Example - Plastic Waste
HS Code - 3915

## Health & COVID-19

| Data | Source | Notes |
|---|---|---|
| Cases, Testing, Healthcare Capacity, Deaths, | https://github.com/MoH-Malaysia/covid19-public | Detailed up to date granular data to state level. |
| Real-time healthcare system capacity | https://www.moh.gov.my/index.php/pages/view/260<br><br>Health Cluster Open Data Portal | Not realtime, annual only. |
| Vaccination | https://github.com/CITF-Malaysia/citf-public | Detailed up to date granular data to state level. |

## International Trade

| Data | Source | Notes |
|---|---|---|
| Import and Exports of Goods and Services | Comtrade<br><br>Malaysian External Trade Statistics | Detailed granular data by HS Code over several years. |

# Case Study 1 - Kini News Lab COVID-19 Tracker

https://newslab.malaysiakini.com/covid-19/en

## Journalist/Organisation

malaysiakini
news and views that matter

Malaysiakini (Aidila Razak (L), Lee Long Hui (R), Wong Kai Hui, Sean Ho, Thiaga Raj Servai, Hazman Hazwan, Syariman Badrulzaman)

*"We managed to create something that was useful to the public at a time when there was so much unknown, and so little information provided by the authorities." - Aidila Razak*

*"The most challenging and frustrating thing about this project was trying to figure out what data needs to be there, how to get the data and to agree as a team."- Aidila Razak*

## Description

This data journalism project by Malaysiakini is a website which includes a dashboard of key statistics regarding the COVID-19 pandemic in Malaysia (at national, state, district and subdistrict level), verified locations affected by COVID-19, cluster information, patient and death information, resources on how to stay safe during the pandemic and rules and regulations of lockdowns and other related information. It is published in English, Bahasa Malaysia and Chinese, with some key parts of the website published in Bengali, Nepali and Burmese to cater to the more than one million migrant worker population who are not English or Bahasa Malaysia literate.

## Challenges

- Data releases inconsistent, health departments decide how and when to release data, inconsistent across states
- Not published in machine readable format, but through websites and social media in the form of infographics and charts.
- The key source of critical data and information for COVID-19

Until detailed data was eventually released by the Ministry of Health on 20th July 2021, KiniLabs COVID-19 tracker was the only central source for cumulative key data related to COVID-19, but also other general information for the public such as movement restrictions and relevant hotlines to call for further information. The site is also the only source of public verified data on local outbreaks needed by others for contact tracing. Malaysian government only provided codenames for clusters, with Malaysiakini having to verify reports on the exact location to update data for the website.

At the time, the data released by the government was not in machine readable format, often in the form of infographics, inconsistent categories, inconsistent numbers reported by different state governments, as well as inconsistency on when data was released. The data, especially state and district level data often had to be collected and entered manually from multiple social media pages such as Facebook.

Daily cases would be updated in daily press briefings for a few weeks, only to suddenly stop for a few days and then resume again at a later date, but with some categories missing. vaccination data would also be released by another Ministry, the Ministry of Science and Technology that managed vaccinations and not by the Ministry of Health.

One of the biggest challenges this project faced was, in collecting the data, became the key source of verified data and information for COVID-19 for the country. The site became more of a public information health portal than a data journalism website. This led to the expectations by the public for it to provide clean, verified and up to date detailed data and information on COVID-19. This eventually led to some parts of the website, such as providing detailed local information by state districts, to be retired.[19] Some of the data and charts are still being maintained by the journalists.

## Methodology and Impact

- Collected data across multiple sources into single Google spreadsheet
- Structured data presented in clear live charts, made it easy for the public to understand and track the COVID-19 situation in Malaysia in a single place.
- Data required for the website and reporting, helped journalists hold the government accountable for consistent, detailed and accurate release of information.
- Data collected, became a valuable source of data for policy makers and researchers

In order to develop the data journalism website with detailed information and visualizations, the KiniLabs team had to manually collect, verify and collaboratively enter data from multiple sources into a Google Docs spreadsheet.

With inconsistent data by the government provided in the form of charts and images, KiniLabs stepped in by making available up to date key indicators and information, in a consistent and clear manner for the public. Tracking and updating the data daily, also led to it being used by Malaysiakini journalists to hold the government accountable when data is not updated or inconsistent.

The data generated was also useful by others including state government and academics. The Selangor state government uses the data collated on locations affected by COVID-19 daily in their monitoring of local outbreaks and roll out prevention and containment strategies. This data was not shared, by the Federal Government[20], which meant it had to rely on data such as that available on request by KiniLabs. The data was also useful for academic researchers.

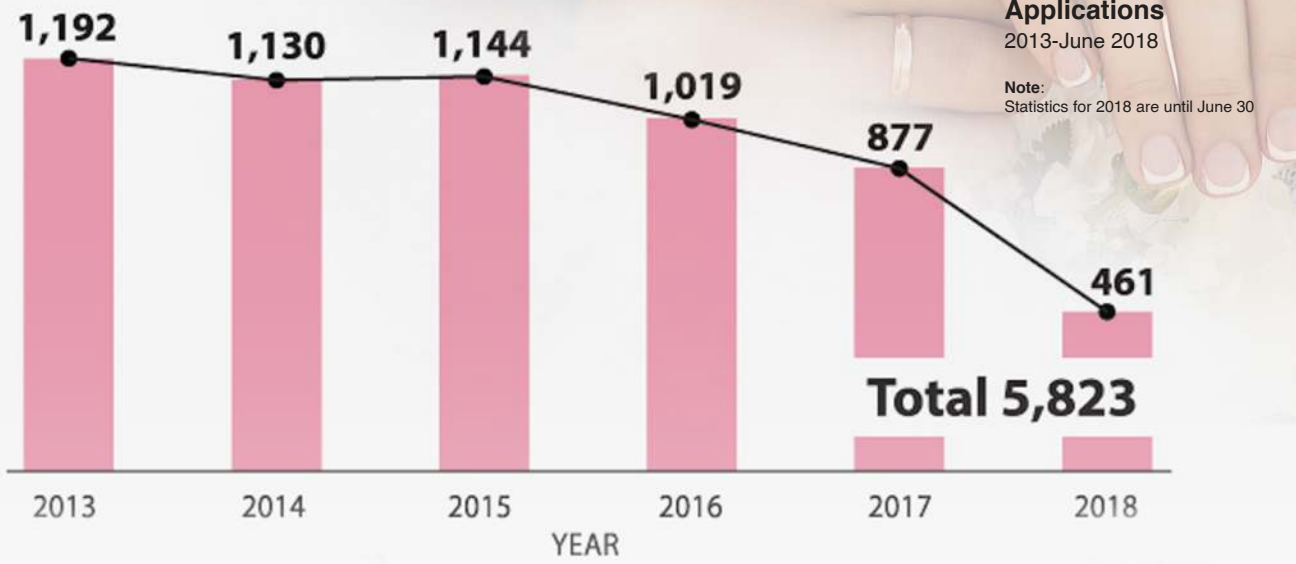
The project has been selected as one of the finalists for the 2021 Sigma Awards for Data Journalism[21].

The Malaysian government eventually started publishing detailed structured[22], open data on COVID-19 on July 20, 2021[23] more than a year after KiniLabs launched COVID-19 Tracker site in March 2020. With the government publishing granular open data and a detailed dashboard[24], the Malaysiakini COVID-19 Tracker site was retired in September 2021 having provided vital information for the public during a health crisis, and affecting change in provision of this information in data in government.

## Sources of Data / Data Generated

- https://covid-19.moh.gov.my/
- https://www.facebook.com/kementeriankesihatanmalaysia/posts/10157672322256237
- Data generated by KiniLabs for this project is available on request to newslab@malaysiakini.com

Source: Department of Shariah Judiciary Malaysia (JKSM) I Graphics: malaymail




# Case Study 2 - Child Marriages in Malaysia

- Three things about: Child marriages in Malaysia
- Ministry: 543 child marriages, including applications, in Malaysia from Jan-Sept 2020

## Journalist/Organisation

malaymail

Ida Lim,
Malay Mail

*"A good way to start a data story is to Google to find out what data is already out there, and possible sources of data. But it is always best to go back to the primary source, and to carry out further checks and verification if needed. For example, a news report that quoted government statistics given in Parliament had some discrepancies, while a check of the Dewan Rakyat's Hansard (the primary source) showed a discrepancy in the total of Malaysia's non-Muslim child marriages (2,725) for 2010-2015. It was after I checked with the relevant deputy minister's office that the accurate total figure of 2,775 was provided and verified." - Ida Lim*

*"Accurate and meaningful official data may prove challenging to obtain, especially in terms of granularity, ease of access and consistency. But it may still be possible to find data that could be helpful to the public." - Ida Lim*

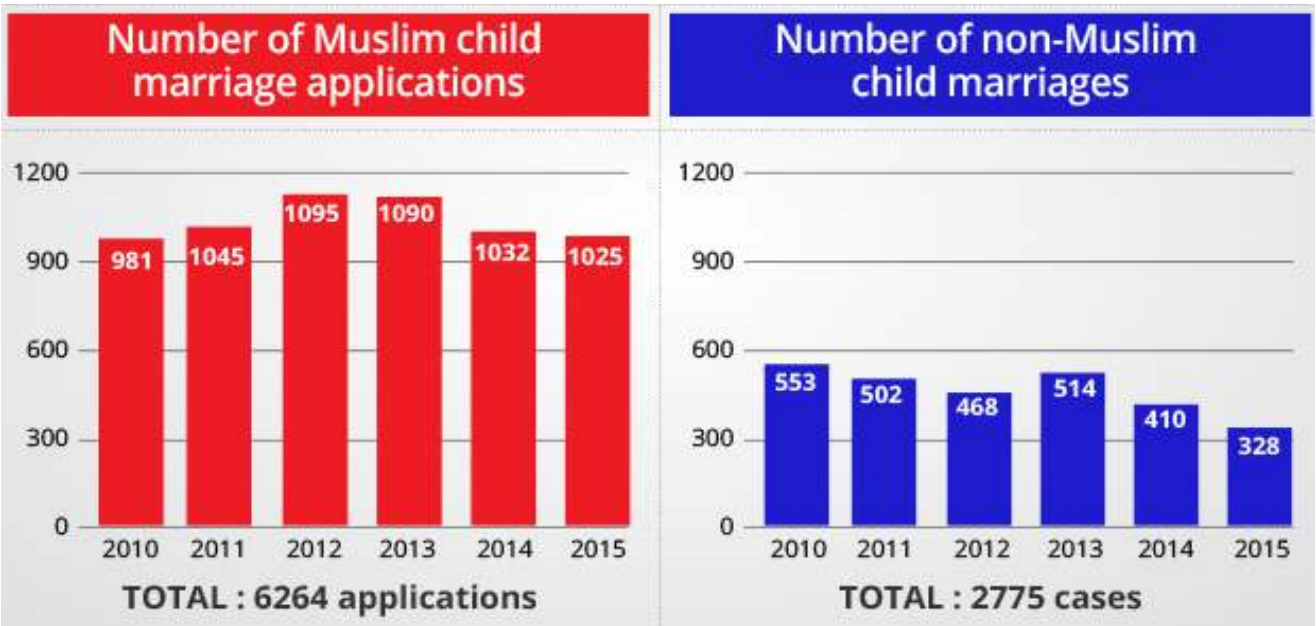
## Description

News reports covering the issue of child marriages in Malaysia, using data for infographics and charts on legal definitions and statistics on child marriages.

## Challenges

- Limited detailed data on child marriages
- Separate legal requirements and sources of data for marriage for Muslims and Non-Muslims

There are continous campaigns to demand an end to child marriage, which is still legal in Malaysia. Marriages in Malaysia come under dual legal systems: Shariah for Muslims and the National Registry Department for non-Muslims. Issues around race and religion are often considered as sensitive in Malaysia, leading to self-censorship on discussions and availability of related data.



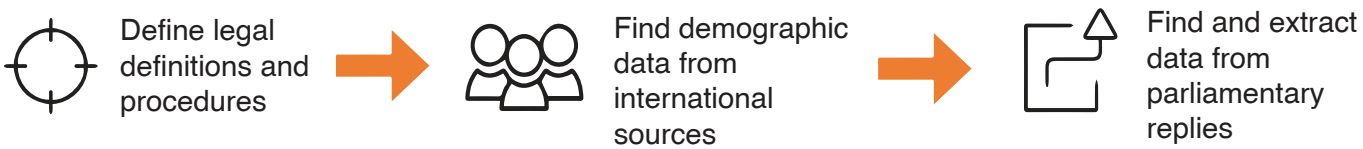

**Graphic: Child Marriages in Malaysia (2010-2015)**
Source: JKSM and NRD figures from Deputy Women Minister Datuk Azizah Mohd Dun's parliamentary reply, May 19, 2016 (Hansard)
Graphic: themalaymailonline.com

Shariah law is further complicated, with different laws applicable for each state in Malaysia, under jurisdiction of state government Islamic agencies.

The lack of granular data made it harder to show a full picture of child marriages in Malaysia. For example, Department of Statistics Malaysia's data for marital status was by a rigid set of age groups (e.g. 10-14, 15-19) instead of by the individual ages (16, 17, 18). This means the data could not be analysed or presented in a way to show child marriages among those aged 18 and below, or aged 16 and below. Inconsistent data is also a problem, as shown by the anomaly in year 2010 for the data on marital status for children in the age group 10-14 where all were recorded as Never Married for this year, when compared to other years.

Therefore, it was a challenge to not only find data, but also to consolidate data on the actual state of child marriages in Malaysia. Lack of data was cited by UNICEF as one of six enabling factors for child marriages in Malaysia.[25]

## Methodology and Techniques




These stories start by first researching clear definitions and legal terms about the data being covered. Clarification on terms then determines where potential sources of Child Marriage data can be found, and where they might be inconsistent. It can then help define a single standard that could be used when consolidating data from different sources, or when it is simply not possible. When data is hard to find, especially for topics not usually covered, informative introductory reporting on the issue, can be just as important as the information gleaned from data found.

With limited open data available, multiple sources of data were used to provide context, the total number of child marriages in Malaysia. This includes census data, data provided by international organisations such as the United Nations Population Division and also parliamentary replies.

### TIPS: Looking for data in Parliamentary Documents

With thousands of answers on Sinar Project's Parliamentary Documents website, try using additional keywords like **jumlah** and **statistik** to have higher chance of finding answer with data or tables

## Sources of Data / Data Generated

- Parliamentary replies https://pardocs.sinarproject.org/documents/2019-march-april-parliamentary-session/oral-questions-soalan-lisan/2019-03-13-parliamentary-replies/par14p2m1-soalan-lisan-36.pdf/view
- 2010 Census data https://www.mycensus.gov.my/index.php/census-product/publication/census-2010
- World Marriage Data (United Nations Population Division)

Photo: From Bukit Awan to Gunung Tahan, by SK Chong

# Case Study 3 - Evaluating Forest Management in Peninsular Malaysia

https://www.macaranga.org/data-story-peninsular-malaysia-forestry/

## Journalist/Organisation

Macaranga

L-R: YH Law, SL Wong for Macaranga

*"What's very challenging to get is an accurate and updated map of forest reserves. I haven't found one. Forestry Deparment doesn't give me any, citing "sulit" (secret) as a reason, and the hardcopies printed in reports cannot be accurately digitised because the outlines are too blurry." - Yao-Hua Law*

*"There are two sources of official data for this (changes in area of forest reserves) - State Gazettes and forestry reports.*

*Frustratingly, the numbers of forest reserve area changes between forestry reports and State Gazettes do not tally. Since such changes are official only after they are published in gazettes, one should use gazettes as the official and primary data source. But to read State Gazettes, you have to pay for digital subscription, and even then you only get those from 2001 onwards.*

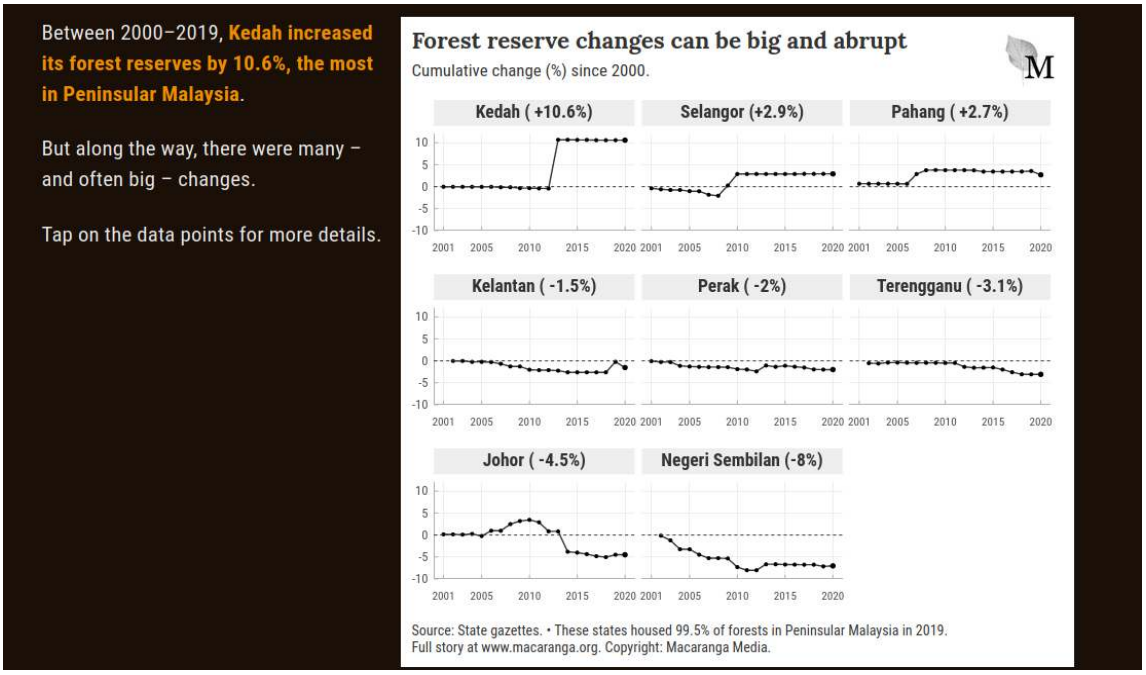
*And State Gazettes can backdate changes in forest reserves - Kelantan's for example has published gazettes in 2010 announcing forest reserve changes effective 2003!*

*To be as accurate as possible, one needs to pay for the gazettes, then check every entry related to forestry in detail."*

*- Yao-Hua Law*

## Description

Interactive survey and data visualizations, to explain the state of Forest Management in Malaysia. Data and documents generated for this project was then published as open data for use by other journalists and researchers.




*Between 2000–2019, Kedah increased its forest reserves by 10.6%, the most in Peninsular Malaysia.*

*But along the way, there were many – and often big – changes.*

## Challenges

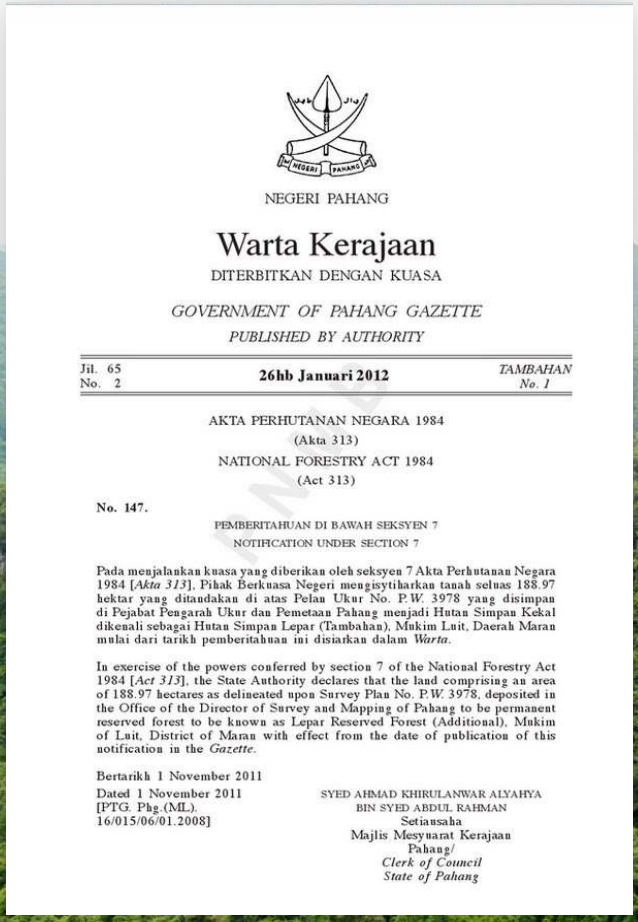
- Lack of open data of forest reserve status and type
- Lack of open data of maps of forest reserves and type

There is limited data available on the status and total area of forest reserves in Malaysia, only aggregate totals by country and state. Degazetted forest reserves may be replaced by being designated elsewhere.

Digital maps of the actual areas covered by the forest reserves were also not available. Finally additional verification was needed on whether designated forest reserve areas are still there or have been logged through the use of satellite imagery.

## Methodology and Techniques

- Explore different types of data for forest governance
- Use satellite data to verify land usage and forest coverage
- Extract data on forest reserve/land usage from State Gazettes
- Find data from civil society



NEGERI PAHANG

Warta Kerajaan

DITERBITKAN DENGAN KUASA

*GOVERNMENT OF PAHANG GAZETTE*

*PUBLISHED BY AUTHORITY*

Jil. 65 No. 2 — **26hb Januari 2012** — *TAMBAHAN No. 1*

AKTA PERHUTANAN NEGARA 1984
(Akta 313)
NATIONAL FORESTRY ACT 1984
(Act 313)

No. 147.

PEMBERITAHUAN DI BAWAH SEKSYEN 7
NOTIFICATION UNDER SECTION 7

Pada menjalankan kuasa yang diberikan oleh seksyen 7 Akta Perhutanan Negara 1984 [*Akta 313*], Pihak Berkuasa Negeri mengisytiharkan tanah seluas 188.97 hektar yang ditandakan di atas Pelan Ukur No. P.W. 3978 yang disimpan di Pejabat Pengarah Ukur dan Pemetaan Pahang menjadi Hutan Simpan Kekal dikenali sebagai Hutan Simpan Lepar (Tambahan), Mukim Luit, Daerah Maran mulai dari tarikh pemberitahuan ini disiarkan dalam *Warta*.

In exercise of the powers conferred by section 7 of the National Forestry Act 1984 [*Act 313*], the State Authority declares that the land comprising an area of 188.97 hectares as delineated upon Survey Plan No. P.W. 3978, deposited in the Office of the Director of Survey and Mapping of Pahang to be permanent reserved forest to be known as Lepar Reserved Forest (Additional), Mukim of Luit, District of Maran with effect from the date of publication of this notification in the *Gazette*.

Bertarikh 1 November 2011
Dated 1 November 2011
[PTG. Phg.(ML). 16/015/06/01.2008]

SYED AHMAD KHIRULANWAR ALYAHYA
BIN SYED ABDUL RAHMAN
Setiausaha
Majlis Mesyuarat Kerajaan
Pahang/
*Clerk of Council*
*State of Pahang*

Macaranga started with first clearly defining several criteria to measure forest management, and then proceeded to try the find data each criteria:

- Area of permanent forest reserve
- Change in area of forest reserves
- Forestry Revenue
- Timber harvest
- Primary forest cover

Data for overall areas of permanent forest reserves and revenue, could be found and extracted from Forestry Department's annual reports. To visualise forest coverage since official map data was not available, Macaranga sourced map data from civil society effort, Hutan Watch that manually created shapefiles from different sources. Meanwhile, primary forest cover map and data were sourced from Global Forest Watch.

In order to calculate actual forest reserve changes, it was discovered that additions and subtractions to designated forest reserves are published in State Gazettes. This information is buried in thousands of public domain PDF documents, behind a paywall. 20 years of forest reserve data was extracted and tabulated to be able to generate the data to track the changes for each state.

Following data journalism best practices, Macaranga published the source documents and data extracted from PDFs of annual reports and gazettes as open data. Not only does it provide readers a way to verify the data-driven story and visualizations, but new and hard to obtain forestry data is now available to other journalists and researchers to generate more data-driven stories.

## Sources of Data / Data Generated

- Forestry Department Peninsular Malaysia - Annual reports
- Hutan Watch
- State Gazettes
- Compendium of Environment Statistics, Department of Statistics Malaysia
- Global Forest Watch
- Data made available from this project https://www.macaranga.org/forestry-data-peninsular-malaysia

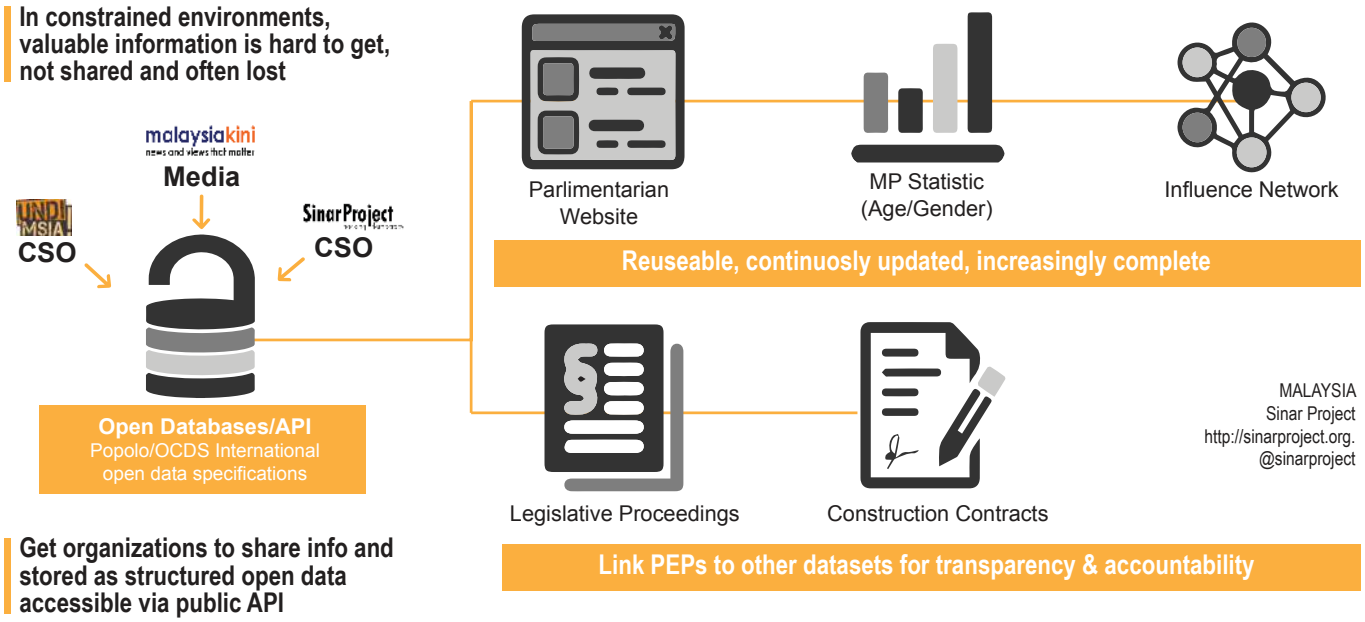
# Case Study 4 - Politically Exposed Persons, Contracts

## Journalist/Organisation

SinarProject hacking democracy

L-R: Khairil Yusof, Ng Swee Meng, Sinar Project

*"In constrained environments, data availability is limited and civil society and journalists have limited resources. Open data standards can be used to build a foundation to collaboration in making hard to obtain data available for use by data journalists." - Khairil Yusof, Sinar Project*




## Description

Use of open data standards to collect data from public information and media reports on politically exposed persons, positions held, relationships and interests to uncover possible corruption and conflicts of interest.

## Challenges

- No data on political integrity such as asset declarations or political financing
- No data on politically exposed persons and relationships
- Limited data on procurement and conflict of interests
- Difficulty finding and combining fragmented data and information from multiple sources
- Data sources are often not available online or taken down

There are daily if not constant media reports on corruption, conflicts of interest and poor governance in Malaysia. While there have been books and research on the extent and involvement of politicians, this has not been captured as data. Concurrent efforts at capturing this network as data was done in this area mainly through: Sinar Project's work with electoral data, construction contracts and media reports[26] with data on politicians and positions in government-linked companies by Terence Gomez, Thirshalar Padmanabhan,

Norfaryanti Kamaruddin, Sunil Bhalla as part of the research for their publication, 'Minister of Finance Incorporated: Ownership and Control of Corporate Malaysia'.[27]

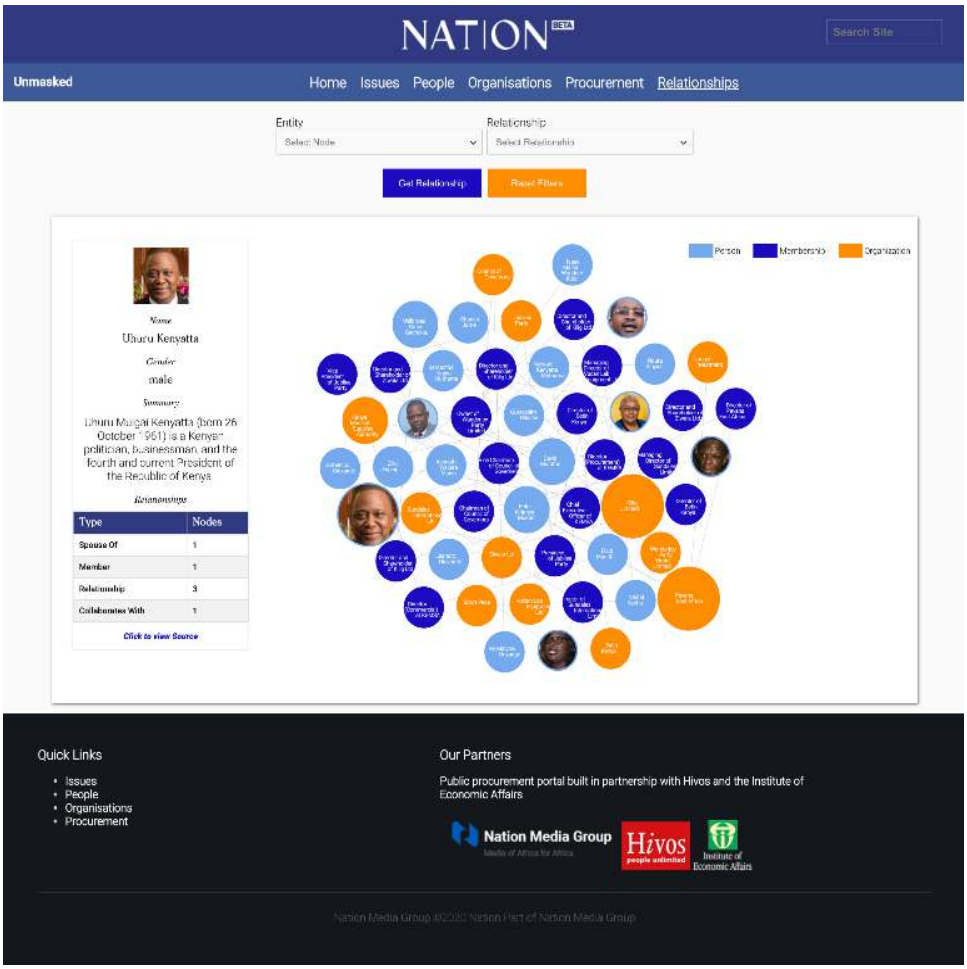
It was found that there was a lot of effort in gathering information by multiple parties, but they were not captured as data in a consistent manner or published. There was a wealth of information, but it was fragmented and often lost and out of date. Additionally, key data points such as exact dates on positions held by politically exposed persons, and verifiable sources were also missing, making it difficult or impossible to use the data to publish a story.

## Methodology and Techniques

- Use of open data standards Popolo-spec, Open contracting, Beneficial Ownership, CoST to structure data to be collected and stored
- Developed open source content management system, to enabled tracking and storage of multiple sources of data as well publication of data
- Work with journalists to ensure editorial integrity of collected data
- Using joined up and linked data to uncover conflicts of interest

In order to find a well-designed and complete data structure to combine fragments of information, open data standards were used as base templates for spreadsheets and databases. These standards have had significant resources and expert input in their development. Repurposing them for data journalism meant that data collected would be well structured, and cover numerous use cases and input from stakeholders internationally, along with technical support of the organisations developing these standards.

The project collaborated with journalists at Malaysiakini, to understand editorial standards, such that the data was also publishable by journalists. This included the ability to store links to multiple sources for each piece of data captured, importance of capturing official contact details for right of reply and clear definitions of relationships. Before relationship data is added and published to link two persons of interest as business partners, it had to be defined clearly, and then sources had to be found to support that editorial definition.




Caption: Application of Politikus approach for Unmasked Kenya, Nation Nation Media

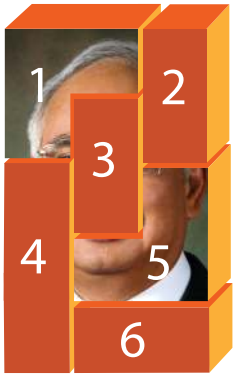
It was found that repurposing a content management system to publish data was more appropriate for data journalism. It allowed the addition of private notes, multiple sources, as well as a publication workflow such that some data accessible by journalists working on the story, were embargoed until the story was published. Originally inspired from Reuters project “Connected China”[28], for media organisations this approach of capturing fragmented information as data, can also increase the value of traditional reporting, by providing additional data for innovative data visualizations. It can provide additional insights for each new story, but also increase the longevity of day-to-day reporting by linking back to them from each new data-driven story. This work was also short-listed for Sigma 2021 Data Journalism Awards[29].

## Sources of Data / Data Generated

- Open Data on Malaysian PEPs, Relationships, Organizations, Beneficial Ownership and Procurement
- Government annual reports, Parliamentary hansards, Securities exchange on Government Documents
- Politikus Source Code
- Unmasked Nation Africa

# Best Practices in Data Journalism in Constrained Environments

## Organize Data Structure Using Open Data Standards and Hypothesis




Structured data standard or template provides a guide for sourcing and generating data

**Use data standards for structure**

**Use online spreadsheet for collaboration**

**Use multiple sources of data to generate data**

From case studies and data journalism training workshops, in Malaysia, journalists have had to find and generate data from multiple sources for stories. This blurs the line between data and investigative journalism, and methods from both are needed in order to source data needed for their stories.

## Example: Water Disruption

### Converting Notices into Structured Data

Caption: Unstructured data from Air Selangor social media notices

Water disruption is a common issue[30] for residents of Klang Valley and granular data is not available. From experience of Kinilab's COVID-19 tracker case study, journalists can apply a similar approach in structuring data-driven stories by collecting and generating data from official announcements into a standard data set. The process of doing so will also help structure common causes, and durations for water cuts.

This will enable journalists to provide much better data- backed insights such as the main causes of outages, average time and days consumers are affected by outage and which areas were affected– beyond simple day-to-day reporting of water outages.

Take note as per example above, that the data needs to be raw and as granular as possible, so one incident notice affecting multiple areas or times, will require a new row entered for each area and time.

| Start Datetime | End Datetime | District | Area | Source | Cause | Source Link |
|---|---|---|---|---|---|---|
| 2021-09-03 17:15 | | Petaling | Taman Serdang Jaya | Air Selangor | Odour for Semenyih River | https://twitter.com/air_selangor/ |
| 2021-09-03 17:15 | | Petaling | Taman Kembangsari | Air Selangor | Odour for Semenyih River | https://twitter.com/air_selangor/ |
| --- | --- | --- | --- | --- | --- | --- |

Caption: Information from Air Selangor notices and other sources converted into structured data

### TIPS: Plan for resources and time to compile data

It takes a lot of effort to manually compile enough granular data from multiple sources, many of which may be in the form of images. Plan ahead to ensure enough resources and adequate time for data compilation before the expected publishing date of the story.

## Open Data Standards - Electoral Data

### Using Data Standards To Improve Data and for Collaboration

KEPUTUSAN PILIHAN RAYA UMUM KE-13 BAGI
BAHAGIAN PILIHAN RAYA PARLIMEN MENGIKUT PARTI-PARTI YANG BERTANDING

| Bahagian Pilihan Raya | Bilangan Pemilih | Bilangan Mengundi | Peratus Undi | BEBAS | BERJASA | BERSAMA | BN | DAP | KITA | PAS | PCM | PKR | SAPP | STAR | SWP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| P.001 - PADANG BESAR | 41,974 | 36,142 | 86.10 | | | | 21,473 | | | 14,047 | | | | | |
| P.002 - KANGAR | 51,207 | 43,431 | 84.80 | | | | 23,343 | | | 19,306 | | | | | |
| P.003 - ARAU | 43,876 | 38,439 | 87.60 | 406 | | | 19,376 | | | 18,005 | | | | | |
| P.004 - LANGKAWI | 37,536 | 32,096 | 85.50 | 180 | | | 21,407 | | | | | 9,546 | | | |
| P.005 - JERLUN | 52,383 | 45,899 | 87.60 | | | | 24,161 | | | 20,891 | | | | | |
| P.006 - KUBANG PASU | 65,550 | 57,296 | 87.40 | | | | 33,334 | | | 22,890 | | | | | |
| P.007 - PADANG TERAP | 41,960 | 37,904 | 90.30 | 243 | | | 20,654 | | | 16,212 | | | | | |
| P.008 - POKOK SENA | 80,714 | 69,524 | 86.10 | | | | 32,263 | | | 36,198 | | | | | |
| P.009 - ALOR STAR | 69,009 | 57,912 | 83.90 | | 3,530 | 257 | 25,491 | | | | | 27,364 | | | |
| P.010 - KUALA KEDAH | 95,328 | 82,253 | 86.30 | | | | 37,923 | | | | | 42,870 | | | |
| P.011 - PENDANG | 70,135 | 62,578 | 89.20 | | | | 32,165 | | | 29,527 | | | | | |
| P.012 - JERAI | 74,410 | 64,778 | 87.10 | | | | 32,429 | | | 31,233 | | | | | |
| P.013 - SIK | 46,786 | 42,077 | 89.90 | | | | 22,084 | | | 19,277 | | | | | |
| P.014 - MERBOK | 85,908 | 74,520 | 86.70 | | | | 38,538 | | | | | 34,416 | | | |
| P.015 - SUNGAI PETANI | 93,176 | 81,024 | 87.00 | 772 | | | 34,646 | | 200 | | | 44,194 | | | |
| P.016 - BALING | 93,168 | 83,109 | 89.20 | | | | 43,504 | | | 38,319 | | | | | |
| P.017 - PADANG SERAI | 74,095 | 64,584 | 87.20 | 669 | 2,630 | | 25,714 | | | | | 34,151 | | | |
| P.018 - KULIM-BANDAR BAHARU | 60,910 | 52,766 | 86.60 | | | | 26,782 | | | | | 24,911 | | | |
| P.019 - TUMPAT | 98,632 | 82,962 | 84.10 | | | | 35,487 | | | 46,191 | | | | | |
| P.020 - PENGKALAN CHEPA | 64,409 | 54,985 | 85.40 | | | | 19,497 | | | 34,617 | | | | | |
| P.021 - KOTA BHARU | 81,268 | 66,277 | 81.60 | 148 | | | 24,650 | | | 40,620 | | | | | |
| P.022 - PASIR MAS | 71,965 | 60,168 | 83.60 | 25,384 | | | | | | 33,431 | | | | | |
| P.023 - RANTAU PANJANG | 52,903 | 41,934 | 79.30 | | | | 17,405 | | | 23,767 | | | | | |
| P.024 - KUBANG KERIAN | 65,390 | 55,108 | 84.30 | | | | 18,769 | | | 35,510 | | | | | |
| P.025 - BACHOK | 81,566 | 71,792 | 88.00 | | | | 35,218 | | | 35,419 | | | | | |

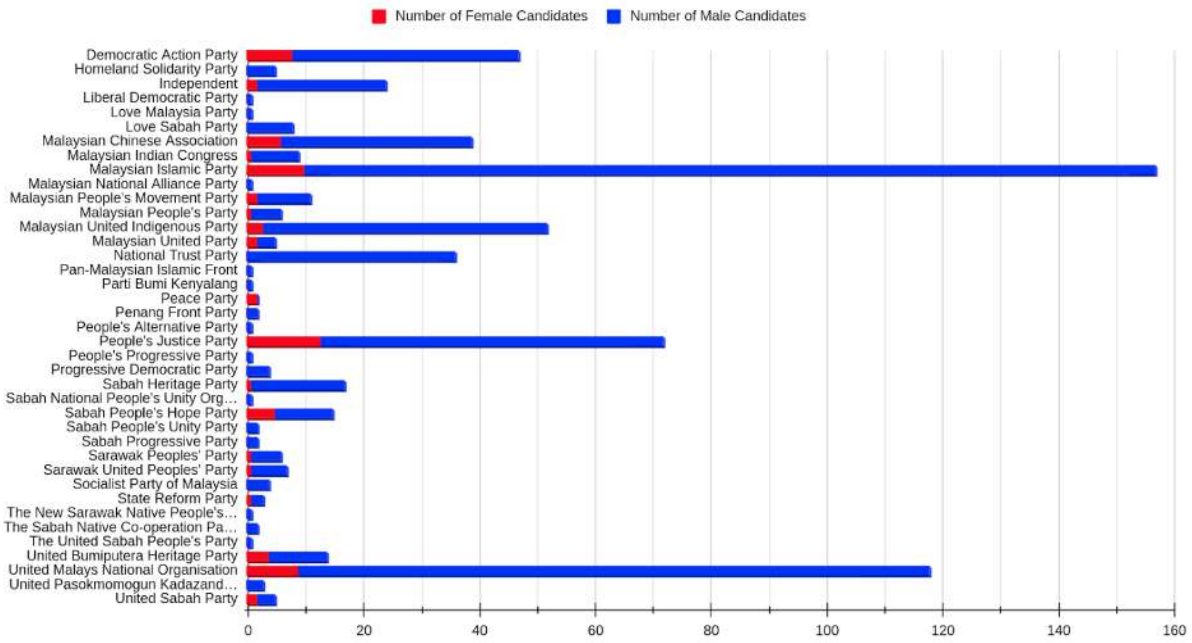
Caption: Official 14th General Elections Results

During elections, official results data often include the bare minimum in terms of medata for an election, often just including the area, the candidate name. The Malaysian system of elections, published data will include only the name of the coalition ticket and not the actual political parties the candidates belong to.[31]

For election stories, journalists need to provide additional insights on reporting on topics such as gender equality, youth or other diversity indicators beyond just results provided.

| bilangan_undi | status | area_identifier | area_name_en | area_name_ms | person_name_en | person_name_ms | gender | post_role_en |
|---|---|---|---|---|---|---|---|---|
| 8,530 | MENANG | P041/N01 | Penaga | Penaga | Mohd Yusni Mat Piah | Mohd Yusni Bin Mat Piah | Male | GE14 Candidate fo |
| 7,398 | | P041/N01 | Penaga | Penaga | Mohd Zain Ahmad | Mohd Zain Bin Ahmad | Male | GE14 Candidate fo |
| 2,986 | | P041/N02 | Bertam | Bertam | Mokhtar Ramly | Mokhtar Bin Ramly | Male | GE14 Candidate fo |
| 6,485 | MENANG | P041/N02 | Bertam | Bertam | Khaliq Mehtab Mohd Isha | Khaliq Mehtab Bin Mohd | Male | GE14 Candidate fo |
| 6,268 | | P041/N02 | Bertam | Bertam | Shariful Azhar Othman | Shariful Azhar Bin Othma | Male | GE14 Candidate fo |
| 7,627 | | P041/N03 | Pinang Tunggal | Pinang Tunggal | Roslan Saidin | Roslan Bin Saidin | Male | GE14 Candidate fo |
| 4,622 | | P041/N03 | Pinang Tunggal | Pinang Tunggal | Bukhori Ghazali | Bukhori Bin Ghazali | Male | GE14 Candidate fo |
| 7,754 | MENANG | P041/N03 | Pinang Tunggal | Pinang Tunggal | Ahmad Zaki Yuddin Abd | Ahmad Zaki Yuddin Bin A | Male | GE14 Candidate fo |
| 5,021 | | P042/N04 | Permatang Berangа | Permatang Bera | Mohd Shariff Omar | Mohd Shariff Bin Omar | Male | GE14 Candidate fo |
| 6,870 | MENANG | P042/N04 | Permatang Berangа | Permatang Bera | Nor Hafizah Othman | Nor Hafizah Binti Othmar | Female | GE14 Candidate fo |
| 6,224 | | P042/N04 | Permatang Berangа | Permatang Bera | Mohd Sobri Saleh | Mohd Sobri Bin Saleh | Male | GE14 Candidate fo |
| 24 | HILANG DEPOSIT | P042/N04 | Permatang Berangа | Permatang Bera | Azman Shah Othman | Azman Shah Bin Othmar | Male | GE14 Candidate fo |
| 5,115 | | P042/N05 | Sungai Dua | Sungai Dua | Yusri Isahak | Yusri Bin Isahak | Male | GE14 Candidate fo |
| 7,314 | MENANG | P042/N05 | Sungai Dua | Sungai Dua | Muhamad Yusoff Mohd N | Muhamad Yusoff Bin Mol | Male | GE14 Candidate fo |
| 5,380 | | P042/N05 | Sungai Dua | Sungai Dua | Zahadi Mohd | Zahadi Bin Mohd | Male | GE14 Candidate fo |
| 7,072 | MENANG | P042/N06 | Telok Ayer Tawar | Telok Ayer Tawa | Mustafa Kamal Ahmad | Mustafa Kamal Bin Ahma | Male | GE14 Candidate fo |
| 3,900 | | P042/N06 | Telok Ayer Tawar | Telok Ayer Tawa | Mohamad Hanif Haron | Mohamad Hanif Bin Haro | Male | GE14 Candidate fo |
| 88 | HILANG DEPOSIT | P042/N06 | Telok Ayer Tawar | Telok Ayer Tawa | Lee Thian Hong | Lee Thian Hong | Male | GE14 Candidate fo |
| 4,869 | | P042/N06 | Telok Ayer Tawar | Telok Ayer Tawa | Zamri Che Ros | Zamri Bin Che Ros | Male | GE14 Candidate fo |
| 2,136 | HILANG DEPOSIT | P043/N07 | Sungai Puyu | Sungai Puyu | Lim Hai Song | Lim Hai Song | Male | GE14 Candidate fo |
| 101 | HILANG DEPOSIT | P043/N07 | Sungai Puyu | Sungai Puyu | Tan Lay Hock | Tan Lay Hock | Male | GE14 Candidate fo |
| 79 | HILANG DEPOSIT | P043/N07 | Sungai Puyu | Sungai Puyu | Neoh Bok Keng | Neoh Bok Keng | Male | GE14 Candidate fo |
| 21,705 | MENANG | P043/N07 | Sungai Puyu | Sungai Puyu | Phee Boon Poh | Phee Boon Poh | Male | GE14 Candidate fo |
| 51 | HILANG DEPOSIT | P043/N07 | Sungai Puyu | Sungai Puyu | Ong Yin Yin | Ong Yin Yin | Female | GE14 Candidate fo |
| 2,898 | | P043/N08 | Bagan Jermal | Bagan Jermal | Ang Chor Keong | Ang Chor Keong | Male | GE14 Candidate fo |
| 106 | HILANG DEPOSIT | P043/N08 | Bagan Jermal | Bagan Jermal | Hari Devyndran Muniswa | Hari Devyndran A/L Mun | Male | GE14 Candidate fo |
| 30 | HILANG DEPOSIT | P043/N08 | Bagan Jermal | Bagan Jermal | Fabian George Albart | Fabian George Albart | Male | GE14 Candidate fo |
| 18,134 | MENANG | P043/N08 | Bagan Jermal | Bagan Jermal | Soon Lip Chee | Soon Lip Chee | Male | GE14 Candidate fo |
| 74 | HILANG DEPOSIT | P043/N08 | Bagan Jermal | Bagan Jermal | Teoh Chai Deng | Teoh Chai Deng | Male | GE14 Candidate fo |
| 3,918 | | P043/N09 | Bagan Dalam | Bagan Dalam | Dhinagaran Jayabalan | Dhinagaran A/L Jayabala | Male | GE14 Candidate fo |
| 10,701 | MENANG | P043/N09 | Bagan Dalam | Bagan Dalam | Satess Muniandy | Satess A/L Muniandy | Male | GE14 Candidate fo |
| 45 | HILANG DEPOSIT | P043/N09 | Bagan Dalam | Bagan Dalam | Teoh Huck Ping | Teoh Huck Ping | Male | GE14 Candidate fo |

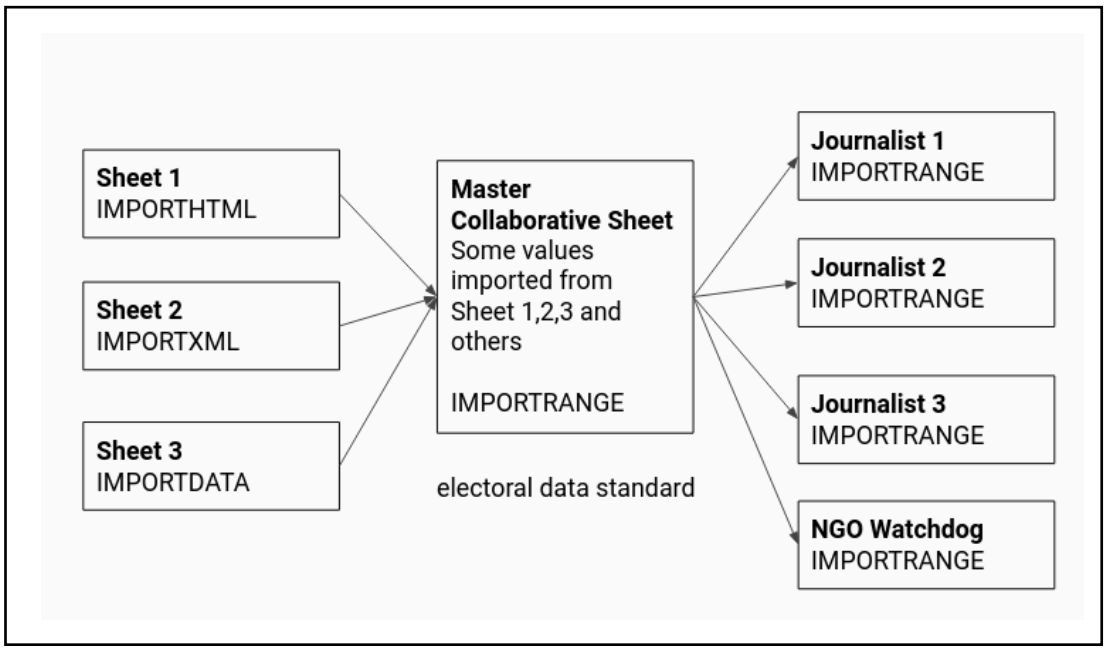
Caption: 14th General Elections Parliamentary Results Data extended with Popolo-spec structure/fields




Caption: Gender breakdown for 14th General Elections Parliamentary Results with extended Popolo-spec data

When data is missing, data standards help us figure out basic columns we should have that we can gather from different sources. Popolo-spec[32] is a good standard for basic fields to capture basic information on people and political parties. When the task of sourcing data is heavy for individual or small media teams, a standard can also help for collaborations between journalists as well as civil society.[33]

Joining data from online spreadsheets also enables collaboration both for sourcing additional data, but also in reusing such data. In fast-changing situations such as elections, where data can sometimes still be in the process of being verified, journalists can import and extend data for their own stories, from the source collaborative spreadsheet and then take snapshots as needed for a story.

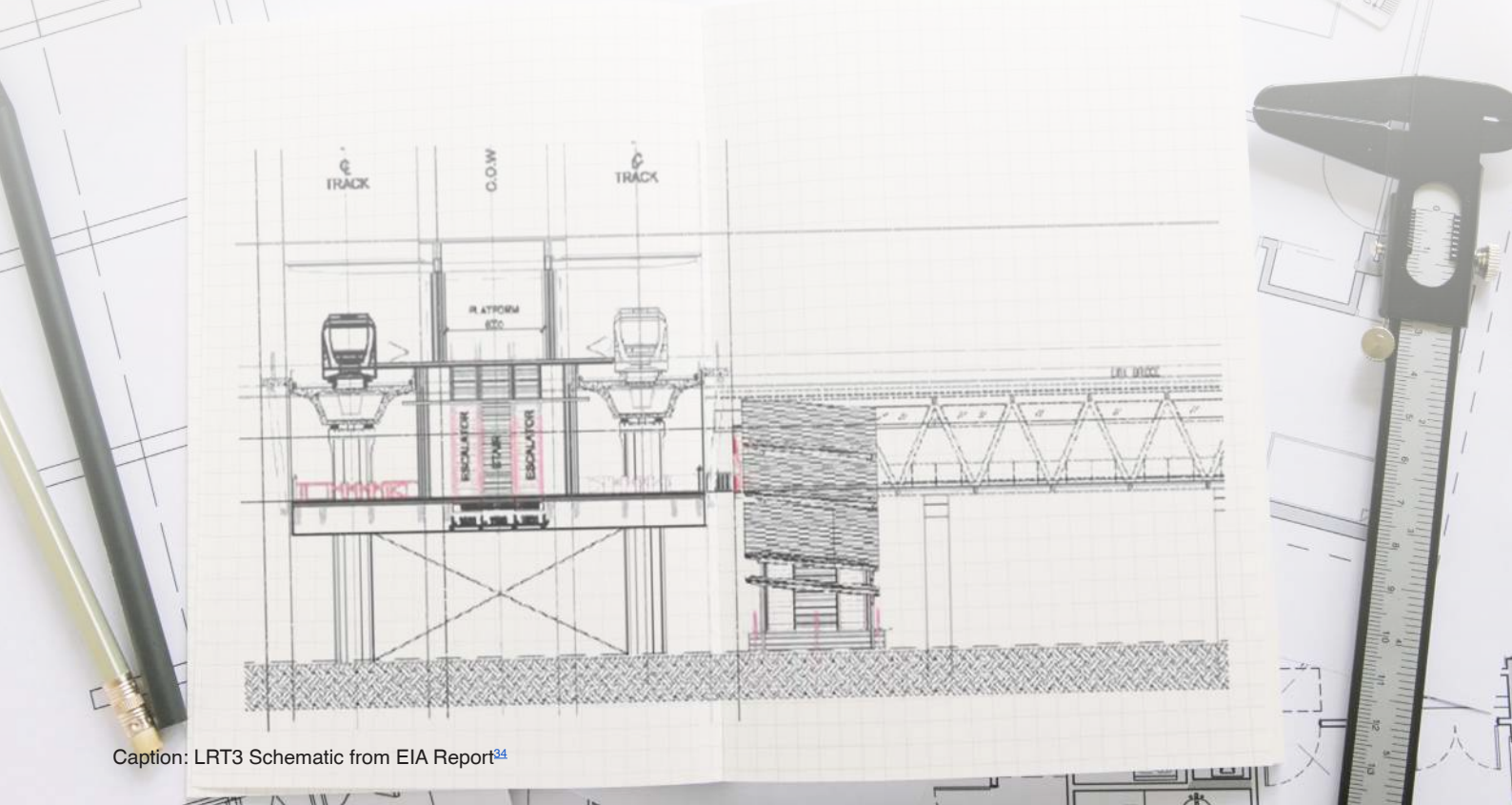



Caption: Example using online Google Sheet's import functions for collaboration in combining and extending multiple sheets.

## Open Data Standards - Infrastructure and Procurement

### Using data standards by sector experts and leveraging government sources

In a constrained environment with limited transparency, using expert reports and standards can be another source of guidance for data journalists. Both in identifying and structuring data sources for data compilation, and in understanding where the issues are to develop a hypothesis. In this example for large scale multi-billion Ringgit infrastructure projects, CoST – the Infrastructure Transparency Initiative (CoST) - Infrastructure Data Standard and Open Contracting Data Standards can be used[35].




Caption: LRT3 Schematic from EIA Report[34]

**CoST Infrastructure Data Standard**

Table 1: Project and Contract Data for proactive disclosure

| Project phase | Project data | Contract phase | Contract data |
|---|---|---|---|
| Last updated | | Procurement | Procuring entity<br>Procuring entity contact details<br>Procurement process<br>Contract type<br>Contract status (current)<br>Number of firms tendering<br>Cost estimate<br>Contract administration entity<br>Contract title<br>Contract firm(s)<br>Contract price<br>Contract scope of work<br>Contract start date<br>Contract duration |
| Project Identification | Project reference number<br>Project owner<br>Sector, subsector<br>Project name<br>Project Location<br>Purpose<br>Project description | | |
| Project Preparation | Project Scope (main output)<br>Environmental impact<br>Land and settlement impact<br>Contact details<br>Funding sources<br>Project Budget<br>Project budget approval date | | |
| Project Completion | Project status (current)<br>Completion cost (projected)<br>Completion date (projected)<br>Scope at completion (projected)<br>Reasons for project changes<br>Reference to audit and evaluation reports | Implementation | Variation to contract price<br>Escalation of contract price<br>Variation to contract duration<br>Variation to contract scope<br>Reasons for price changes<br>Reasons for scope and duration changes |

Using CoST IDS to organize spreadsheet

In previous examples such as election results and water supply disruptions, the data structures are relatively simple with a few sources of information. For large complex opaque procurement projects that span multiple years and layers or procedures, such as infrastructure, creating a spreadsheet to source and structure data becomes a daunting task. Finding and referring to specialist standards can help guide journalists structure the spreadsheet. CoST IDS standard here can be used as reference.

Using the same standard, but looking at references for project documentation can also be utilised by the journalists to find or inquire about official documents that are needed as sources for data. Data journalists can also use reporting by other journalists to piece together project details.

**TIPS: Use Spreadsheet Cell Comments to Track Sources**

Use the cell comment feature of spreadsheets to store notes on sources for each piece of data

Standards can also help guide journalists develop a hypothesis-driven story, by understanding the common issues for a sector or industry. It can then provide a common framework or situations where this problem can occur. The Open Contracting Data Standard (OCDS)[37] provides consistent stages in procurement

Table 2: Project and Contract Information for disclosure upon request

| Project information | Contract information |
|---|---|
| **Identification and Preparation** | **Procurement** |
| Multi-year programme & Budget | Contract officials and roles |
| Project brief or Feasibility study | Procurement method |
| Environmental and social impact assessment | Tender documents |
| Resettlement and compensation plan | Tender evaluation results |
| Project officials and roles | Project design report |
| Financial agreement | **Contract** |
| Procurement plan | Contract agreement and conditions |
| Project approval decision | Registration and ownership of firms |
| | Specifications and drawings |
| **Completion** | **Implementation** |
| Implementation progress reports | List of variations, changes, amendments |
| Budget amendment decision | List of escalation approvals |
| Project completion report | Quality assurance reports |
| Project evaluation report | Disbursement records or payment certificates |
| Technical audit reports | Contract amendments |
| Financial audit reports | |

CoST IDS project information

| PLANNING | TENDER | AWARD | CONTRACT | IMPLEMENTATION |
|---|---|---|---|---|
| Key planning documents not provided | Non-public bid opening or single bidder only | High number of contract awards to one bidder | Large difference between contract award and final contract amount | Modifying the contract after it's been awarded, on line item requirements |
| Eligibility criteria for deciding which companies can bid for a contract set too narrowly | Use of direct awards/ exceptions/ emergency procedures | Company has no history in providing service or product | Conflicts of interest | Turning a blind eye on shoddy implementation |
| | Vague description of supply terms | Cover pricing: Colluding to drive up prices artificially | Supplier receives multiple single source contracts | Change orders to increase prize substantially (or multiple by a smaller amount) |
| | Issue of tender at an inconvenient time | Winning bid is at a substantially lower bid price than competitors or too close to estimate | Final prize is higher than industry average | Payment without delivery of service |
| | Short notice to bidders | Similarity in supplier addresses | | |
| | | Bidder that has never bid previously wins tender | | |

Caption: Open Contracting Data Standard Stages and Problems[36]

that are being adopted by leading economies. By understanding where common problems lies within public procurement, journalists can then quickly develop a hypothesis, along with the data sources they will need to find.

Table 2: Project and Contract Information for disclosure upon request

| Project information | Contract information |
|---|---|
| **Identification and Preparation** | **Procurement** |
| Multi-year programme & Budget | Contract officials and roles |
| Project brief or Feasibility study | Procurement method |
| Environmental and social impact assessment | Tender documents |
| Resettlement and compensation plan | Tender evaluation results |
| Project officials and roles | Project design report |
| Financial agreement | **Contract** |
| Procurement plan | Contract agreement and conditions |
| Project approval decision | Registration and ownership of firms |
| | Specifications and drawings |
| **Completion** | **Implementation** |
| Implementation progress reports | List of variations, changes, amendments |
| Budget amendment decision | List of escalation approvals |
| Project completion report | Quality assurance reports |
| Project evaluation report | Disbursement records or payment certificates |
| Technical audit reports | Contract amendments |
| Financial audit reports | |

### TIPS: Use International Data Standards and Reports for Data Structure

Data is often incomplete, and requires compiling from multiple sources. Look for data and reports published by international bodies such as the World Health Organisation, Food and Agriculture Organisation, International Labour Organisation, UNDP and others to help design your data collection spreadsheet columns and categories for combining data from multiple sources.

## Searching Government Documents and Websites

Due to the difficulty in finding data, the existing sources of data listed may not cover the data needed or may no longer be accessible in the future. The following methods will help journalists find the data they need.

### Using Search Engine Modifiers

When data cannot be found easily on a website, use the following operators to help find the data you need.

<search keywords> site:*.gov.my

and to narrow it down to specific file types and websites

<search keywords> site:data.gov.my filetype:xls

This is useful even for CKAN data portals like data.gov.my because search engines will also return results within the files and not just the descriptions.

**TIPS: Find official search terms and use both Bahasa and English**

## Government Documents

When machine readable formats such as CSV and XLS cannot be found, then data can also be found in government documents.

### Parliamentary and State Legislative Documents

Parliamentary and state legislative documents provide a wealth of information and data[38] that are of public interest, in the form of replies to questions by elected representatives, select committees as well as documents that are submitted to these bodies and often also uploaded to respective websites. Additionally the answers provided in parliament will also reference the source government agencies, from which to get additional data from through enquiries or respective websites.

- parliment.gov.my
- pardocs.sinarproject.org

## Government Reports

A lot of statistics are published as non-machine readable data in various government reports. Reports that provide a lot of information and data include but not limited to:

- Annual Reports
- Auditor General Reports
- Financial Statements
- Environmental Impact Assessment Reports
- Circulars

In addition to web search, many of these documents are also searchable at govdocs.sinarproject.org

**TIPS: Look for indirect data**

Search for direct terms may not return any results. Terms and conditions of government agreements are often unavailable under the Official Secrets Act. Try searching for indirect data.

Example:
Toll and other Concessions Agreements.
The agreement may not be available, but data such as government expenditure may be found in Parliamentary Documents or in Auditor General Financial Statements for Federal and State.

## Securities Commissions

The Malaysian government holds a substantial stake in a lot of key public listed companies, many providing essential services. The annual reports, financial statements and offer documents are publicly available and provide additional information in the sectors and markets these companies are involved in.

In addition the corporate websites, public disclosures and reports of public listed companies can be search using search engines with the following parameters:

keywords site:disclosure.bursamalaysia.com

**TIPS: Extract Tables from PDFs with Tabula**

Data is often found as tables in reports and documents. Quickly extract and export them with Tabula software https://tabula.technology

## Investigative Journalism Resources

Finding data in Malaysia is a challenging process. Many of the techniques shared require innovative and creative ways of finding data, which are often the same methods used by investigative journalists. In addition to the methods covered in this report, many more investigative methods can be found at the Global Investigative Journalism Network (GIJN) website, which will be essential for Malaysian data journalists.

### GIJN Reporting Tips and Tools

https://helpdesk.gijn.org/support/solutions/articles/14000036502-reporting-tips-and-tools

Given potential risks faced by journalists in Malaysia in finding and reporting data, it is recommended that they familiarize themselves with the safety tips from the resources, including protecting sources (see whistleblowing[39]), along with some security basics[40].

## Biodata of Author

Khairil Yusof is an investigative data journalist and researcher working on applying innovative methods of open data and standards, for transparency and anti-corruption. In addition to developing and supporting journalists on investigative methods on corruption, he is also an experienced digital security trainer.

He is the founder of Sinar Project, an organisation that collates patchy government statistics and turns them into usable data for the public and journalists.

He has over 15 years of cross-practice and cross-sector experience implementing programmes for UN agencies, governments and civil society in Asia-Pacific including digital rights, human rights, labour rights, and environment.

# Endnotes

[1] Malaysian Administrative Modernisation and Management Planning Unit (2015), *Pelaksanaan Data Terbuka Sektor Awam*, https://dasar.mampu.gov.my/search-g/download-file/25/7f821c650c868d025fb5351d7d45d001

[2] PEKELILING KEMAJUAN PENTADBIRAN AWAM BILANGAN 2 TAHUN 2021, https://dasar.mampu.gov.my/search-g/download-file/259/e2aebef35549528275d8b47af883b937

[3] Portal Data Terbuka (2021), https://web.archive.org/web/20210419183103/https://www.data.gov.my/

[4] Open Data Barometer 4th Edition (2016), https://opendatabarometer.org/4thedition/?_year=2016&indicator=ODB

[5] Open Knowledge Foundation(2015), Global Open Data Index. http://2015.index.okfn.org/place/

[6] Ahmad Ashraf Ahmad Shahrudin (2021), Open Government Data in Malaysia: Landscape, Challenges and Aspirations, *Khazanah Research Institute Discussion Paper* http://www.krinstitute.org/Discussion_Papers-@-Open_Government_Data_in_Malaysia-;_Landscape,_Challenges_and_Aspirations.aspx

[7] Iqbal Harith Liang (2018), *Freedom of Information in Malaysia*, University of Malaya Law Review, https://www.umlawreview.com/lex-in-breve/freedom-of-information-in-malaysia3869206

[8] Tan Karr Wei (2019, 21 April), Local council papers to remain under OSA, *The Star* https://www.thestar.com.my/news/community/2009/04/21/local-council-papers-to-remain-under-osa/

[9] IFEX (2007), Blogger arrested under Official Secrets Act, another under investigation; symptomatic of clampdown on online expression, says SEAPA, *IFEX.* https://ifex.org/blogger-arrested-under-official-secrets-act-another-under-investigation-symptomatic-of-clampdown-on-online-expression-says-seapa/

[10] Article 19 (2015), Malaysia: Repeal Section 203A of the Penal Code https://www.article19.org/resources/malaysia-repeal-section-203a-penal-code/

[11] Alyaa Alhadjri (2020, 28 June), Cops clarify 'OSA probe' report, says editor-in-chief questioned under Penal Code, *Malaysiakini* https://www.malaysiakini.com/news/532080

[12] Intellectual Property Corporation of Malaysia (1987), *Copyright Act 1987*, https://www.myipo.gov.my/wp-content/uploads/2016/12/Copyright-Act-1987-Act-332.pdf

[13] Ahmad Ashraf Ahmad Shahrudin (2021), Open Government Data in Malaysia: Landscape, Challenges and Aspirations, *Khazanah Research Institute Discussion Paper*

[14] World Bank (2017), Open data readiness assessment : Malaysia https://documents.worldbank.org/en/publication/documents-reports/documentdetail/529011495523087262/open-data-readiness-assessment-malaysia

[15] Reporters Without Borders (2021), 2021 World Press Freedom Index, https://rsf.org/en/ranking/2021

[16] Kit Yong Ng (2004). *From The First Line To The Byline: Malaysian Journalists' Learning In Practice Under The Power Of Media Ownership* [PHD Thesis, The University of Georgia] https://getd.libs.uga.edu/pdfs/ng_kit_y_200405_phd.pdf

[17] Reporters Without Borders (2020, 7 August). RSF denounces Malaysia's harassment of Al Jazeera journalists. https://rsf.org/en/news/rsf-denounces-malaysias-harassment-al-jazeera-journalists

[18] World Bank (2017), Open data readiness assessment : Malaysia

[19] Aidila Razak [@aidilarazak]. (2020, 11 December). https://twitter.com/aidilarazak/status/1337215043834761217

[20] Aidila Razak & Hariz Mohd (2020, 19 October). Selangor hamstrung as Covid-19 data-sharing halted, *Malaysiakini* https://www.malaysiakini.com/news/547174

[21] Sigma Awards (2020), The Kini News Lab Covid-19 Tracker, https://sigmaawards.org/the-kini-news-lab-covid-19-tracker/

[22] Aidila Razak (2020, 7 July). How pushing for transparency helped Malaysians navigate COVID19, *DW Akademie*. https://p.dw.com/p/3f4He

[23] Ministry of Health (2021). Official data on the COVID-19 epidemic in Malaysia https://github.com/MoH-Malaysia/covid19-public

[24] Ministry of Health (2021). CovidNow in Malaysia https://covidnow.moh.gov.my

[25] UNICEF Malaysia (2021), Advocacy Brief: Towards Ending Child Marriage in Malaysia https://www.unicef.org/malaysia/media/1801/file/Child_marriage_brief_factsheet_%28ENG%29.pdf

[26] Michael Canares, Khairil Yusof, Ng Swee Meng (2017), Collaborating For Open Source: Building and Open Database on Politically Exposed Persons in Malaysia: A Case Study http://webfoundation.org/docs/2017/08/RP-Collaboration-For-Open-Data-082017.pdf

[27] Terence Gomez with Thirshalar Padmanabhan, Norfaryanti Kamaruddin, Sunil Bhalla and Fikri Fisal (2017), Minister of Finance Incorporated: Ownership and Control of Corporate Malaysia

[28] Connected China (2014), Reuters http://china.fathom.info/

[29] Sigma Awards (2020), Unmasked by Nation Media Group, https://sigmaawards.org/unmasked-by-nation-media-group/

[30] Jerry Choong (2020, 21 October), A history of water cuts in Selangor this year, Malay Mail. https://www.malaymail.com/news/malaysia/2020/10/21/a-history-of-water-cuts-in-selangor-this-year/1914721

[31] Ben Tan (2018, 16 April), Dr M: All Pakatan Harapan parties to use PKR logo for GE14, Malay Mail https://www.malaymail.com/news/malaysia/2020/10/21/a-history-of-water-cuts-in-selangor-this-year/1914721

[32] Popolo International Open Government Data Specifications (2013), https://www.popoloproject.com/

[33] Khairil Yusof, (2021), *Working With Data, Southeast Asia Data Journalism Training on Election Issues*, https://docs.google.com/presentation/d/1xGcKbBpT93lrza2YwqUXHbWP5x4kyJM3G7iVfLy90nw/edit#slide=id.g89d9d5ffe4_0_77

[34] ERE Consulting Group (2015), Proposed Light Rail Transit Line 3 froM Bandar Utama to Johan Setia, Detailed Environmental Impact Assessment https://govdocs.sinarproject.org/documents/ministry-of-natural-resources-and-environment/eia-reports/proposed-light-rail-transit-line-3-from-bandar-utama-to-johan-setia/201502181557310-1-declaration-form-deia-team.pdf/view

[35] CoST – the Infrastructure Transparency Initiative (2017), The CoST Infrastructure Data Standard, https://infrastructuretransparency.org/our-approach/disclosure/

[36] Open Contracting (2019) Using data to uncover corruption in public procurement

[37] Open Contracting Data Standard (2015), Open Contracting Partnership, https://standard.open-contracting.org/latest/en/

[38] Sinar Project (2018), Sinar Project Parliamentary Answers as Source for Data Demand and Government Oversight https://sinarproject.org/open-parliament/updates/parliamentary-answers-as-source-for-data-demand-and-government-oversight

[39] Global Investigative Journalism Network, *Working with Whistleblowers* https://gijn.org/whistleblowing/

[40] Global Investigative Journalism Network, *Safety and Security*, https://helpdesk.gijn.org/support/solutions/articles/14000036509-safety-and-security

Disclaimer:
*This publication was produced with the financial support of the European Union. Its contents are the sole responsibility of Internews and do not necessarily reflect the views of the European Union.*