COLLABORATING FOR OPEN DATA

BUILDING AN OPEN DATABASE ON POLITICALLY EXPOSED PERSONS IN MALAYSIA: A CASE STUDY

RESEARCH PAPER

WORLD WIDE WEB FOUNDATION

Sinar Project
Empowering Malaysian Citizens

TABLE OF CONTENTS

· ·

· ·

[1] *Regional Research Manager for Asia, Open Data Lab Jakarta World, Wide Web Foundation*

[2] *Coordinator, Sinar Project*

[3] *Senior Software Engineer, Sinar Project*

1. Introduction

In the third edition of the Open Data Barometer (ODB), Malaysia is placed 51st out of 92 countries and eighth among 12 countries in East Asia and the Pacific. Malaysia's position in the ODB, just slightly above China and Vietnam, speaks to the fact that open data is fairly new to the country, which has been beleaguered recently with grand corruption cases involving key leaders in government. It was only a year ago that the chief secretary of state announced that the government of Malaysia would introduce open government principles into its bureaucracy, provide a set of open data guidelines to government agencies, and select national open data champions to pursue the proactive disclosure of information[1]. As part of this initiative, the government has launched an open data portal[2] which contains roughly 1000 datasets (as of May 2016).

However, the portal does not contain any information strengthening accountability among elected and senior public officials. According to Transparency International Malaysia, it is rare for officials to be held accountable for any public issue[3] in the country. Anti-corruption experts in Malaysia have argued that it can often take up to several years before mismanagement or corruption issues in public projects are raised[4]. It then becomes difficult to trace who was responsible, especially if positions have been shuffled around or the companies and departments have been closed down. The auditor general's audit reports[5] as well as press reports rarely note who the heads of the responsible departments or companies were.

In Malaysia, it has been reported that political parties, politicians, senior public officials and businesses are highly connected, with political parties being involved in business, and with politicians and senior public officials holding posts in multiple government-linked companies (including investment companies - GIC)[6]. With only one coalition government in power since independence, many politicians (elected and unelected) and ex-senior public officials continue to hold important positions which affect the way democratic institutions function. In this environment, public information on potential conflicts of interest, beneficial ownership, and political accountability are often hard to come by. There are no legal frameworks – such as public declaration of assets and interests, or freedom of information laws – for gaining access to such information.[7] Furthermore there are restrictions on civil society and press freedoms which constrain citizens' ability to make use of open data that the government has committed to release[8].

---

[1] http://ksn.jpm.gov.my/home.php?frontpage/speech/detail/2011

[2] http://www.data.gov.my/

[3] http://transparency.org.my/wp-content/uploads/2016/03/TI-M-Newsletter-Jan-2016-1.pdf

[4] http://www.abc.net.au/4corners/stories/2013/09/30/3857148.htm

[5] https://www.audit.gov.my/index.php?option=com_content&view=-category&id=81&Itemid=299&lang=en

[6] https://m.malaysiakini.com/opinions/39516

[7] http://www.icnl.org/research/monitor/malaysia.html

[8] http://www.amnestyusa.org/research/reports/critical-repression-freedom-of-expression-under-attack-in-malaysia

2. The Initial Building Blocks

The current political landscape and the lack of government information inspired the Sinar Project to explore the power of open technology and applications to systematically make important information public and more accessible to the Malaysian people. Through the Sinar Project, a group of civic tech activists bonded to work together to promote open governance and defend digital rights of citizens. The project aims to improve governance and encourage greater citizen involvement in the public affairs of the nation by making the Malaysian Government more open, transparent, and accountable.

The Sinar Project previously tried to solve the information scarcity problem by collecting and organizing information from different sources, but lack of time and resources meant that the project ended up rewriting common components for different projects multiple times. The Popit[9] component from the Poplus Project, an international collaboration of civic technology developers, was suggested by MySociety, a UK-based non-governmental organisation (NGO) that builds civic engagement tools. Popit is a database built with an application programming interface (API) for storing information on politicians using Popolo[10], which is a set of open governance data specification standards. Popit was originally only meant to store information on elected representatives, but due to the flexibility of the Popolo governance standards, the Sinar Project and a few other civil society groups[11] shaped Popit for local needs, where data on politicians and private and public sector are explored.

Table 1. Initial outputs from the database on politicians and company ownership

Initial projects	Details
MyMP – parliamentarian website http://mymp.undimsia.com/en	The details on all 222 current MPs on this website were sourced from election candidate data, past surveys and current MP surveys from three different partners. The shared open data and API from Popit allowed the site to be built and information updated automatically. Additionally, automated reports were written using open data on MPs, and showed the potential for reusing open data. Script to generate Twitter List of MPs: https://github.com/Sinar/sinar.myreps/blob/master/docs/Malaysian%20MP%20Social%20Media%20Accounts.ipynb iPython Notebook on age and gender of MPs: https://github.com/Sinar/sinar.myreps/blob/master/docs/Malaysian%20MP%20Statistics.ipynb
MyStars – matches politically exposed persons (PEPs) with construction contracts http://mystars.sinarproject.org	At the 16th International Anti-Corruption Conference, a small team of developers from around the world developed a concept application that matched politicians from the Sinar Project's Popit database via an API so that the public could help verify and explore potential linkages between PEPs and construction contracts.
neo4J network graphs	As an experimental exercise, the Popit database (which contained relationships in the form of memberships and parent/child organisations) was imported into the neo4j network graph database. The dynamic graphs provided insights into complicated linkages between politicians and government-owned entities, most notably the recent 1MDB scandal involving state-owned companies and the allegation that billions of Ringgit in funds were transferred into the prime minister's bank accounts. [t-1] This output showed the possibilities and flexibility of shared open databases, which had not been the original purpose of using the open data on politicians (see Figure 1 below).

9   http://popit.poplus.org

10   http://www.popoloproject.com

11   See for example Cargografias Uruguay at http://www.cargografias.org/cargografias#/name-

t-1   http://edition.cnn.com/2015/09/04/asia/malaysia-1mdb-scandal-explainer

Basic personal information on most parliamentary representatives in Malaysia is not easily available. In 2013, crowdsourcing and surveys undertaken as part of the MyMP project resulted in only partial information being recorded on 60 out of 222 members of parliament (MPs). As a result, the Sinar Project learned that in order to build profiles detailing the past employment and posts held by public representatives, this information must be pieced together from various sources. This reinforces the necessity of working on a single, comprehensive and reusable database so that services and tools with a public API can be developed according to widely-accepted standards.

After six months (January to July 2015) of research and development, and with support from the Southeast Asia Technology and Transparency Initiative,[12] the Sinar Project acquired information on over 4000 politicians/persons of interest, 200 organisations/committees, and 900 government posts, as well as mostly complete basic data on all 222 MPs in the current parliamentary session. The Sinar Project was able to show the promise of this approach by creating consumer applications and research notes from this single, combined database within a short time frame and a limited budget. Table 1,

below, provides a summary of the initial outputs of this approach.

A central open database on elected officials supports parliamentary openness by systematically linking officials to the outputs of legislative assemblies and all government agencies, as well as by holding elected officials and organisations accountable to the public. When linked to legislative outputs such as parliamentary responses or government actions on specific issues or target groups (such as gender issues or child rights), open data would empower these groups by allowing them to see what policies and measures have been taken – and their effectiveness – by matching them to research from organisations investigating these issues.

Corruption and lack of transparency are major issues in Malaysia. The current approach of building connections between seemingly disparate and incomplete datasets via a central open, flexible and standardised governance database is starting to show how good governance can benefit from using open data and tools to expose the interests of politically exposed persons (PEPs) and thus hold them accountable.

neo4j graph generated from Sinar Project public popit database https://sinar-malaysia.popit.mysociety.org/api
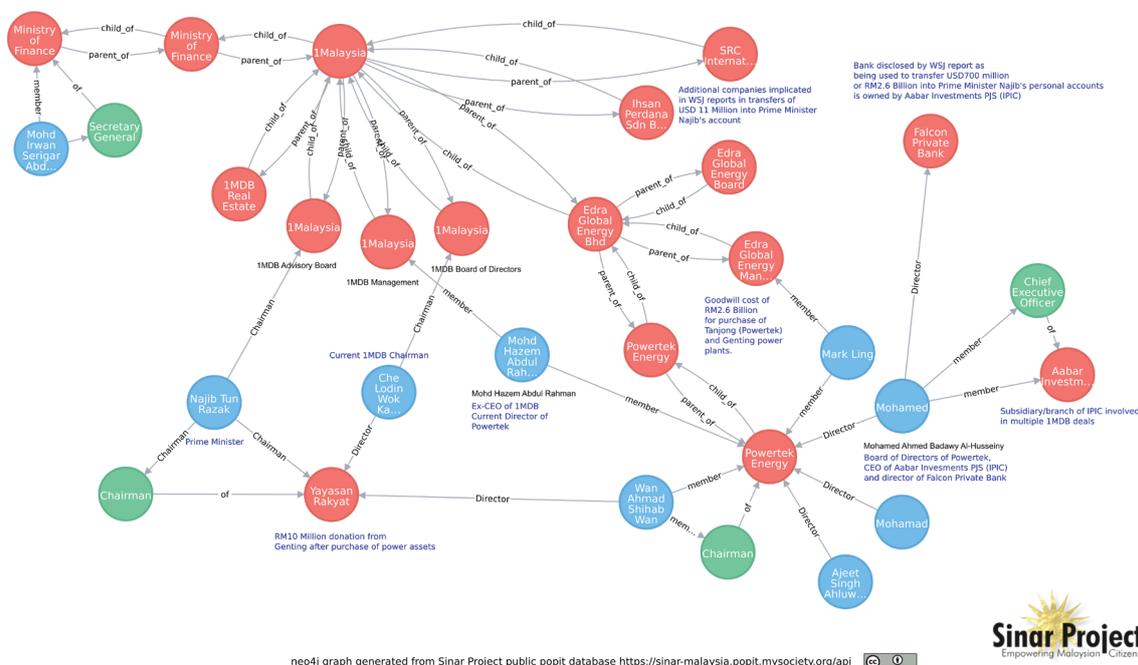
Figure 1. Persons and organisations connected to 1MDB, IPIC and Falcon Private Bank

3. The Need for More (and Better) Data

The Sinar Project's initial work surfaced the need for more and better data that can provide context to the linkages between political interests, business ownership and the behaviour of public officials (both elected and appointed). Advocacy groups and civil society organisations (CSOs) in Malaysia are also clamouring for more data that would help them hold government officials accountable. For example, there is a need to map out the contacts of all senior officials (not just representatives/ policy-makers) because budgets at both national and local government level are often managed by department heads. There is a need to understand how members of legislative assemblies vote on particular issues that affect business interests. There is also a need to understand the relationship between budget and spending data and the key socio-economic indicators meant to influence the allocation of funds.

Unfortunately, though, there is scant information in the government portal or on ministry websites that can answer these questions. There is instead a wealth of unstructured information from various sources. Thus, this action research sought to explore ways to:

- Gather data from different platforms, sources, stakeholders, or data-collecting or data-aggregating agents;

- Build a model for collaboration and data-sharing in order to piece together different data and create a compelling narrative; and

- Create a shared database that users can access for different purposes.

To initiate this process, the team used the process framework illustrated in Figure 2.

In this case, the Sinar Project (i) defined the data needed; (ii) explored potentially good sources of data; (iii) experimented with using open data tools to extract only valuable information; and (iv) tested the data's usability for different users, particularly, CSOs.


Identifying Data Needs → **Exploring Data Sources** → **Extracting Data** → **Testing Usability**
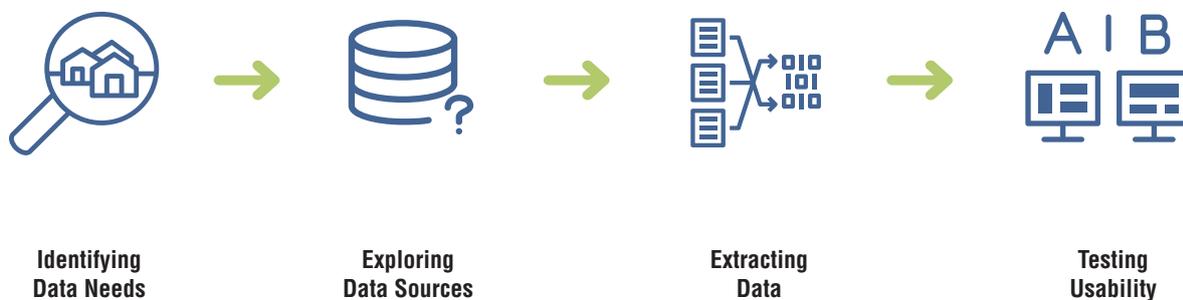
Figure 2. Action research process

4. Sources, Stakeholders and Processes

The dichotomy of data suppliers and data users is a hypothetical distinction – a single stakeholder can be both a user and a supplier of data (Van Schalkwyk et al. 2016). In undertaking this project, data users, in their desire to build an openly available dataset, collected data from various sources, aggregated them and then passed them to others as an open resource.

As a result of this process, different documents – from government, companies and the media – became the basis for the aggregated dataset built to respond to the different data needs of users. Table 2 below summarises the results of this process.

Apparently, several organisations are interested in building databases that would satisfy various data and information needs. For this project, Sinar Project partnered with a number of organisations and individuals which provided the data that can be stored using Popolo standards. These organisations are indicated in Table 3 below.

As indicated in Table 3, there is strong interest among different stakeholders from within government, CSOs, and private citizens to build a central database that would

Table 2. Information needs and sources

Data needed	Example	Identified source	Process	Partner
Various data related to a state, represented by an MP	What information is available on contractors awarded particular contracts for a given year?	Parliamentary documents (replies) and the public accounts committee	Textual analysis of parliamentary replies and the scraping of relevant data (See Annex 1)	Parliamentarian asking specific and particular questions
		Construction Industry Development Board Contracts Database		
Data on private interests held by public officials	Who are the board of directors, shareholders, officers of a particular company?	Company documents that can be requested from the Company Commission[t-2]	Textual analysis of company data and scraping of relevant data (See Annex 2)	
		Corporate websites	Saving archived copies and scraping online information (See Annex 3)	
		Company ownership and directors from Securities Commission Malaysia and archives of finance sections of newspapers	Textual analysis of company data and databases of public listed companies	Academics
Data on ongoing investigations, conflicts of interest, and other related matters	Who is implicated in a particular corruption scandal?	Newspaper articles, investigative reports	Saving archived copies, undertaking textual analysis, scraping and backward confirmations of relevant data (See Annex 4)	Media companies, investigative reporters, bloggers
Data on various construction projects and the ownership of companies engaging in these projects	Who owns the companies that were granted government construction contracts?	Construction Industry Development Board database (http://smb.cidb.gov.my/directory/contractors)	Textual analysis of company data and scraping of relevant data (see http://mystars.sinarproject.org)	

[t-2] https://www.ssm.com.my/en

respond to the data needs of different stakeholders. The Sinar Project, in this case, (i) facilitated intermediation between and among these different stakeholders; (ii) provided the data skills to aggregate and link data; and (iii) standardised data curation, publication and storage. Building the data skills of different users was also critical, as it enabled them to engage with and make use of the data in their current work. In this case, data collaboration is possible among various organisations with varying aims but with common data needs.

Table 2 shows that two main processes were undertaken by different partners (mentioned in Table 3), namely, textual analysis and the scraping of relevant data. As most of the data were unstructured and several of the datasets were saved as pdf files, various partners, including the Sinar Project, did a textual analysis of the data and then scraped only those data relevant to the information needs articulated by different users. For example, researchers from the university did a textual analysis of data from publicly-listed companies and then scraped the most relevant data, such as the names of company owners and directors.

Table 3. Partner profiles

Name of partner	Profile
MCCHR	Malaysian Centre for Constitutionalism and Human Rights (MCCHR) is a non-profit organisation based in Kuala Lumpur with the mission of promoting active democratic participation and human rights.
Coalition for Good Governance, Integrity, Accountability and Transparency	Coalition of civil society organisations including Transparency International Malaysia, Institute for Democracy and Economic Affairs (IDEAS), Centre to Combat Corruption and Cronyism (C4) and the Sinar Project that promotes and advocates for transparency, integrity, accountability and good governance in Malaysia and raises awareness of issues pertaining to corruption.
Prof. Dr. Terence Gomez, Faculty of Economics & Administration, Universiti Malaya	Edmund Terence Gomez specialises in state-market relations and the linkages between ethnicity, politics and capital development. Author of Malaysia's Political Economy: Politics, Patronage and Profits.
Tindak Malaysia	The organisation concerns itself with electoral reforms and voter education. Tindak has several projects that collect and use data for fairer and better electoral boundaries. http://www.tindakmalaysia.org/
Office of Ong Kian Ming, Serdang, MP	Ong Kian Ming is the MP for Serdang http://ongkianming.com

5. Data Collaboration? Data Collaboratives?

Individual and organisational interests and incentives drove this process – within a context of data scarcity. The intersection of different interests – specifically, the need to know – made possible the collaborative building of datasets in order to reveal patterns, generate insights and create knowledge. In data (or open data) literature, the concept of a data collaborative, or data collaboration, has been gaining traction, largely because of the huge need to obtain data on different critical sectors (e.g. agriculture and climate) and the corresponding scarcity of such available, or openly available, data.

Open data collaboration is defined in several ways. For the Global Open Data on Agriculture and Nutrition (GODAN) initiative, it provides a space where 'researchers and practitioners can share, discover, analyse and discuss open data'.[13] In a related argument, Verhulst and Sangokoya (2015) argue for what they call a 'collaborative', which they refer to as 'a new form of collaboration, beyond the public–private partnership model, in which participants from different sectors—including private companies, research institutions and government agencies—can exchange data to help solve public problems'.[14] Looking into these definitions, still at an early stage of conceptualisation, there are at least three major common elements.

The first element is the 'felt' need for quality data, which can be: (i) the need to know (i.e. the transparency argument); (ii) the need to solve a problem (i.e. the effectiveness argument); (iii) the drive to make things efficient (i.e. the efficiency argument); or (iv) the desire to create new things (i.e. the innovation argument). This list, of course, is not exhaustive, and may include several other arguments defined by context and driven by personal or organisational motivation. In the case of the Sinar Project, it was the perceived lingering opacity of the Malaysian government and the desire of several stakeholders to strip away the alleged veil of secrecy involving government finances that prompted the different actors to act on whatever available data they could access.

The second element is the multiplicity of actors – often coming from different sectors – working cooperatively to realise their individual motivations, which are nevertheless intersecting at data. The hypothetical construct of supplier

and user, in this case, gets blurred. Each actor becomes a user and a supplier at the same time. The manner in which each actor behaves, in so far as the data asset is concerned, is different to the others. As the GODAN definition suggests, some share, some discuss, some analyse, and others read, appreciate and learn. The sectoral categorisation also gets blurred because there is no longer the construct of public and private, for profit or not-for-profit; partners are identified by their roles rather than by the nature of their affiliation or by their responsibilities. As in the case of the Sinar Project, the minister becomes a user of the data that the collaboration was able to produce, though in essence, the government should have been the one disclosing the data.

The third and final element is the notion of collaboration. Interdependency becomes more prominent because each piece of the puzzle one is holding is important for weaving a more coherent story across actors and themes. The need for better data, within a context of data scarcity, creates the impetus for this collaboration to occur. Without this element of collaboration, achieving the individual or organisational goals or satisfying the need (element one) of the different actors (element two) could not have occurred. It is interesting that this collaboration is not governed by any formal agreement or by written rules, but by a social contract that binds the relationships among actors, who, by their own volition, opt in, and who, because of this very characteristic, can freely opt out.

It can be argued, then, that the glue that binds the actors in this case is their common need (element one). Satisfying this need is critical for the sustainability of data collaboration. As long as this optimum goal is not yet achieved, actors may continue to work together. It does not mean though that the scarcity of data needs to prevail to sustain the relationship, because the opposite can also happen. It can also be the case that frustration over a long process of co-creation can leave people tired and disillusioned.

How is this collaboration created and maintained? In the case of the Sinar Project, which arguably is still in an early stage of collaboration, the process is facilitated largely by the Sinar Project acting as a convener. But it can also be that, in other instances, the process will naturally evolve via a common need and be conditioned by constant interactions among actors. In the case of the Sinar Project, the process underwent four phases (see Figure 2) – data needs identification, data source mapping, data extraction and data use.

13   http://www.godan.info/blog-posts/prototyping-new-data-collaboration-platform-sustainable-development

14   https://medium.com/@sverhulst/data-collaboratives-exchanging-data-to-improve-people-s-lives-d0fcfc1bdd9a#.lreg17jon

6. Initial Results and Implications

One of the results of this collaboration was the Sinar Project's recently launched portal which hosts datasets from the project as well as from the Malaysian CSOs mentioned in Table 3. To allow for inter-linkages between datasets, the Sinar Project continued the development of Popit 2.0 and the database API to serve its contextual uses, including attempts to link the database of parliamentarians with parliamentary documents and application allowing users to view PEPs and their relationships with other individuals and corporate entities.

The Sinar Project facilitated a series of sessions with civic hackers not only to promote the use of the data, but also to build a community able to discuss different issues concerning government transparency and corporate ownership. They also conducted introductory training sessions on open data using the database as a concrete data example. Collaboration with different partners – from academia, civil society organisations, hackers and even government officials – resulted in greater awareness regarding the importance of the data that the Sinar Project had collated. As a result, a few cases of the database's use have been documented to date.

Prof. Terrence Gomez, an academic from Universiti Malaya, mapped PEPs with government-linked companies with assistance from the Sinar Project and used the project's data. Parliamentarians who were identified as champions for open data, and who supported freedom of information, could then use the Sinar Project's parliamentary documents site as a resource. Outside Malaysia, the Sinar Project assisted Myanmar's Multingual Parliamentary Monitoring website with api.openhluttaw.org using the Popit API.

Admittedly, there is a wide gap between data availability that the Sinar Project tries to positively act on and the capacity of users to leverage the data. Narrowing this gap entails investments in the capacity of prospective users to make meaningful use of the data. Consistent with the findings of other studies on open data, data supply does not automatically lead to use (Canares and Shekhar 2016) as there are many barriers to overcome to make this happen (Davies et al. 2014). Some experts have argued that it requires several things, namely, (i) context (i.e. the receptiveness of government); (ii) data supply (i.e. the quality of the data); and (iii) intermediary skills (i.e. the interest and capacity of civil society.) The ability of users to understand data, and the capacity of intermediaries such as the Sinar Project to make expert language understandable and meaningful is vital to the project's success. The Sinar Project, in this case, performs a critical intermediary role while at the same time is challenged to ensure its tech-heavy language and capacities can be translated into concepts and operational constructs that other intermediaries can use.

Among other things, the Sinar Project was able to effectively illustrate the point that, in countries and environments with a lack of government-provided open data, stakeholders can still take advantage of the work of open data standards and tools with data contributed by CSOs. Public databases using open data standards such as Popolo provide a standardised way of collaboratively producing data from different sources and making the data accessible and usable by different stakeholders. The promise of this process has been appreciated by other stakeholders outside Malaysia. In Myanmar, the Sinar Project assisted a local CSO to build a multilingual parliamentary monitoring website.[15]

---

[15]   http://openhluttaw.info/en_US/

7. Concluding Remarks

This case study shows that in contexts where government data are not available (or when available, are not open), it is possible to implement open data initiatives from below – through the very people who are in need of and are interested in the affairs and workings of its government. While far from being perfect, the experience of the Sinar Project confirms an earlier conceptualisation by Verhulst and Sangokoya (2015) of the conditions that accelerate data collaboratives, albeit in a different way. In their conceptualisation, the collaborative does not operate in a constrained context like Malaysia – where there seems to be an intentional refusal by government to provide information, or to provide government information in a way that is hard to understand such that critical stories and narratives are difficult to uncover.

The experience of the Sinar Project yields some important conclusions:

1. **The data demand and supply must match**. Understanding the demand means understanding the need for data: knowing what the users' interest is and comparing this to what data is currently available. In the case of the Sinar Project, the needed data required intensive work from a number of stakeholders to build a database from different sources, sometimes from scratch.

2. **Data experimentation is essential.** There is a need to use different tools and methods to collaboratively produce and use data. In this case, the Sinar Project used Popolo standards.

3. **Training the stakeholders involved to produce and use data is critical.** Training is a necessary step to ensure collaborative projects can deliver actual products and outputs.

4. **Documentation of the impact of the collaboration is needed.** This helps to provide evidence about the usefulness of what was done as part of the collaboration, as well the collaboration process itself.

Admittedly, the Sinar Project is still struggling with the third and fourth components. However, it has already built the foundation to ensure the sustainability of the collaboration that has been undertaken. Given the challenges of the Malaysian context in terms of data and civil society space, this is an impressive accomplishment.

For more information on this project, visit our website at labs.webfoundation.org or get in touch with us directly at **contact@webfoundation.org**. Other related media are also available in the resources section of our website.

↘ If you want to learn more about the project, email us at **contact@webfoundation.org**

↘ If you want to look at other open data projects, see **www.webfoundation.org**

↘ If you want to give it a try and implement the approach – with or without us – let's talk!

REFERENCES

Canares M and Shekhar S (2016). Open data and subnational governments: Lessons from developing countries. Journal of Community Informatics12(2). Available at http://ci-journal.net/index.php/ciej/article/view/1260 [accessed 22 August 2016]

Davies T(2014)Open Data in Developing Countries: Emerging Insights from Phase 1.Available at http://opendataresearch.org/content/2014/704/open-data-developing-countries-emerging-insights-phase-i [accessed on 20 March 2016]

GODAN (2016) Global Open Data for Agriculture and Nutrition: Success Stories Issue 1.Available at http://www.godan.info/documents/godan-success-stories-issue-1 [accessed on 15 June 2016]

Van Schalkwyk F., Canares M, Chattapadhyay S & Andrason A (2016). Open data intermediaries in developing countries. Journal of Community Informatics, Special Issue on Open Data

Verhulst S & Sangokoya D (2015) Data collaboratives: Exchanging data to improve people's lives. Blog entry available at https://medium.com/@sverhulst/data-collaboratives-exchanging-data-to-improve-people-s-lives-d0fcfc1bdd9a#.ouv07dojn [accessed on 12 August 2016]

ANNEXES

Annex 1. Information from Parliamentary Inquiries
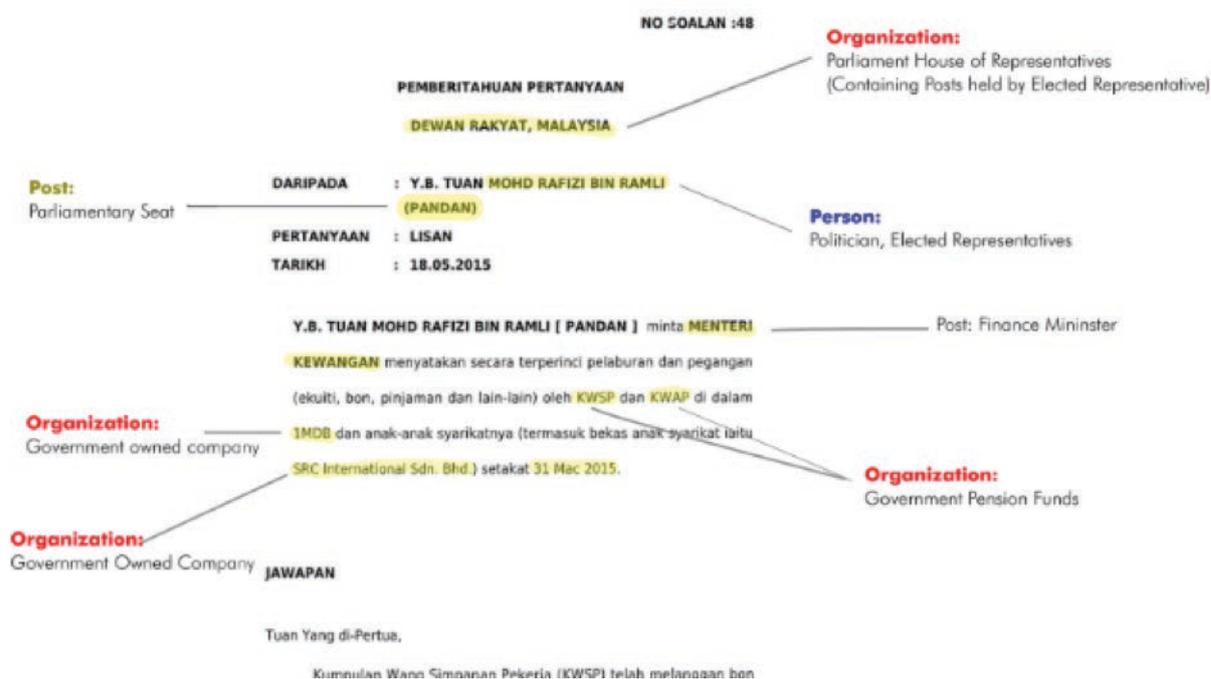
Annex 2. Company Data from Company Commission

Annex 3. Company Data from Corporate Websites

World Wide Web Foundation

webfoundation.org
@webfoundation
contact@webfoundation.org

WORLD WIDE WEB
FOUNDATION