KHAZANAH
RESEARCH
INSTITUTE

VIEWS 2/22 | 28 FEBRUARY 2022

Different Policy Priorities for Different Data Types

Rachel Gong

Introduction

Big data analysis and predictive algorithms are a large and unseen part of everyday lives, both online and offline. Hidden biases in data are turning out to have a larger predictive effect than anticipated, for example, gender discrimination in hiring software that filters CVs and job applications[1].

Data may have been first commodified by the private sector, but the public sector is quickly finding data increasingly necessary to deliver services. The use of data to develop digital or data intelligence requires good governance to ensure equitable, non-exploitative use of data for public interest.

Views are short opinion pieces by the author(s) to encourage the exchange of ideas on current issues. They may not necessarily represent the official views of KRI. All errors remain the authors' own.

This view was prepared by Dr. Rachel Gong, a researcher from the Khazanah Research Institute (KRI) with assistance from Yeung Yeu-Mynn, an intern at KRI. The author is grateful for valuable comments from Ashraf Shaharudin and Nazihah Noor.

This article draws from material presented by the author at a webinar hosted by the RSIS Centre of Excellence for National Security on "Surveillance and Privacy during COVID-19 and the Next Normal" in April 2021.

Author's email address: **rachel.gong @krinstitute.org**

Information on Khazanah Research Institute publications and digital products can be found at www.KRInstitute.org.

Cover photo by Towfiqu barbhuiya on Unsplash.

[1] Chen et al. (2018); Dastin (2018)

Risks of abuse and misuse notwithstanding, the collection and use of anonymised personal data can be useful in many policy areas. For example, location data can reveal travel patterns indicating where traffic flow and public transportation should be improved.

However, considering that data can be—and has been—misused, there are several critical data governance questions we need to ask in this rapidly digitalising world: Who owns our data? Who has access to our data? How is our data being protected? What are the laws and regulations in place governing the access, use and storage of our data? This article aims to highlight several key considerations in developing data policy, which is a complex subject well beyond this article's scope that warrants more attention.

Different types of data…

There are many types of data, which can be categorised in several ways. There exist public and private data, both of which can include administrative data, operational data and transactional data. There is also data that is generated by other data, such as analytical and statistical data, more commonly known as research data[2].

The public-private distinction classifies data by funding source, i.e. whether the funds used to collect, store and analyse the data are derived from public or private sources. Public-private partnerships may exist where data ownership is determined by agreement. This article highlights two other dimensions of data that overlap along the public-private line: personal data and behavioural surplus (see Figure 1 below).

Malaysia's Personal Data Protection Act 2010 (Act 709) defines personal data as "any information… that relates directly or indirectly to a data subject, who is identified or identifiable from that information…[3]" such as one's address or bank account number.

Behavioural surplus is a term popularised by Shoshana Zuboff in her work on surveillance capitalism[4]. It is data generated as a by-product from a particular behaviour, analagous to the body language of digital behaviour. For example, imagine an individual searches for a restaurant address. The search term is a piece of transactional data, but the time of the search, the type of device used, the location of the individual performing the search are all behavioural surplus, which is especially beneficial in predictive algorithms developed by technology companies and has largely gone unregulated by governments.

Public census data has traditionally been heavily relied on for public planning, for example, to count children living in a specific area to determine the number of schools and playgrounds needed in that area. Now, technology companies are discovering that searches for things such as places to buy school uniforms or school supplies can yield similar conclusions while providing additional information. For example, search analyses can reveal how much parents consider

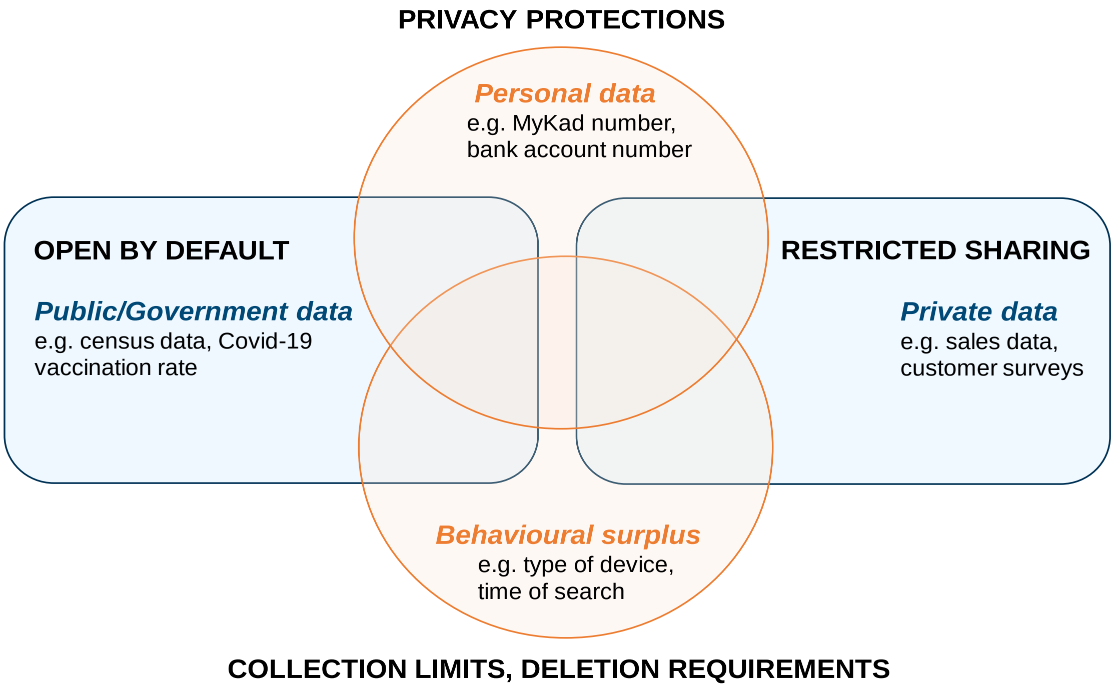
[2] UK National Data Strategy (2020)

[3] Personal Data Protection Act 2010 (2010)

[4] For more information on surveillance capitalism, see Zuboff (2019)'s book, "The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power".

spending on their children's education and where additional public aid and support might need to be provided. As such, behavioural surplus containing personal data collected using private funds can have great value for the public good.

Figure 1: Summary of the different data types and policy priorities with examples


PRIVACY PROTECTIONS
Personal data
e.g. MyKad number, bank account number
OPEN BY DEFAULT
Public/Government data
e.g. census data, Covid-19 vaccination rate
RESTRICTED SHARING
Private data
e.g. sales data, customer surveys
Behavioural surplus
e.g. type of device, time of search
COLLECTION LIMITS, DELETION REQUIREMENTS

Note: Coloured italics indicate data types (blue quadrilaterals indicate data funding distinctions; orange circles indicate data generation distinctions); bold capitals indicate policy priorities.
Source: Author's visualisation

…require different policy priorities

Different types of data require different policy priorities (see Figure 1 above). For example, making public data open by default should be a priority to facilitate data-driven decision-making. Public data are collected and administered using public funds, and thus should be accessible to the public for research and analysis[5].

Private data, on the other hand, could justifiably be deemed proprietary, hence the priority should be focused on allowing firms to use their marketing data to improve their products and services. At the same time, restrictions on companies sharing customers' personal data are needed to protect individual privacy and to ensure that data monopolies do not develop.

Many people are familiar with privacy concerns encircling personal data. In Malaysia, for example, the government is exempt from regulations governing personal data[6]. This exemption

[5] For a discussion of open government data in Malaysia, see chapter 5 of Networked Nation (KRI 2021).

[6] Furthermore, while the Personal Data Protection Act 2010 (2010) protects personal data related to commercial transactions, protection of personal data collected for non-commercial purposes, for example, health records or hiring records, is subject to interpretation.

warrants a review, especially considering how contact tracing data and medical records are being collected in the MySejahtera app and centralised in a public sector database.

Behavioural surplus currently lies predominantly in the hands of the private sector, so in addition to limiting data sharing, policymakers should also prioritise limiting the amount of data that can be collected and how long it can be kept.

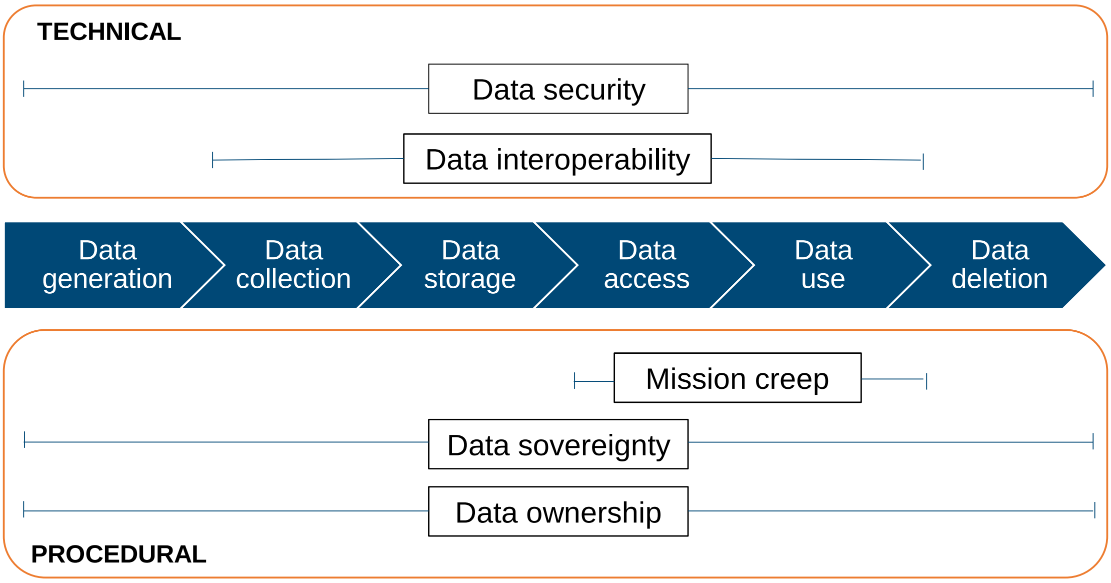
There is a tendency to rely on one overarching policy to address all the needs of data governance[7]. The General Data Protection Regulation was one of the earliest umbrella data regulations passed in 2016. However, data policies resemble most other policies—no one size fits all. The European Union subsequently developed other laws to govern different types of data[8]. Understanding the data value chain can help policymakers decide what data governance strategies to prioritise.

Data governance is needed all along the data value chain

At its most basic, the data value chain begins at data generation, followed by data collection, then data storage, data access, data use and eventually data deletion. As mentioned earlier, data use can generate new data, creating a loop in the chain. Data governance is needed all along the data value chain.

Broadly speaking, there are at least five aspects of data governance that can be divided into two groups: technical aspects and procedural aspects (see overview in Figure 2).

Figure 2: An overview of data governance along the data value chain



Source: Author's visualisation

[7] That said, a national data framework can be overarching and standardised to facilitate interoperability and data sharing with specific regulations and guidelines for different types and uses of data.
[8] For example, the Open Data Directive, the Data Governance Act, the Data Act and the Digital Services Act.

Technical aspects

Two key technical aspects of data governance are data security and data interoperability. Data interoperability is essentially enabling data sharing by design, facilitating a data commons that allows people to combine data for different uses.

Data security is required not only for dynamic transactions, for example in encrypting payment details, but also for static storage, ensuring that databases are secure. The weakest link in cybersecurity is generally not the system itself but the human user[9].

Procedural aspects

Procedural aspects of data governance include mission creep, data sovereignty and data ownership, all of which are the subject of much discussion and debate.

Mission creep occurs when data is used beyond its original intentions. On one hand, it can improve efficiency and lead to innovations, such as personalised shopping recommendations. However, it can also be a breach of trust with discriminatory effects e.g. standardised test scores being used in an algorithm to select scholarship students disregards students who did not sit for those tests, resulting in a form of data invisibility.

Data sovereignty refers to the idea that stored data should be subject to the laws and customary practices of the host country[10]. This is especially relevant for countries that are employing cloud computing on servers that may be located in a different country. Should citizen data be stored outside the country and subject to another country's regulations or lack thereof?

It is one thing to say public data should be stored locally, but what about backups of emails and documents idling on commercial servers? How should cross-border data flows and digital trade be managed? Some countries, such as China and India, heavily regulate foreign apps, but the EU and the United Kingdom remain committed to international cross-border data flows as part of digital trade[11].

In terms of data ownership, questions remain over rights to the use of data, especially for economic purposes. Does data ownership lie with the individual (as the data source), the platform as the point of collection or the storage provider, or whomever is the primary user generating value? For example, does data generated using MySejahtera belong to the user, to the company who made the app, or to the Ministry of Health?

Should individual data owners have the right to trade their data? How might this disadvantage those who lack digital or socio-economic resources? Given that people can pay to remove tracking cookies and other surveillance tools, already disadvantaged groups may have to sacrifice their data and their privacy for digital access, thus creating a new type of digital divide.

[9] Gratian et al. (2018)
[10] Taylor (2020)
[11] Cory and Dascoli (2021)

Conclusion

In summary, data governance is needed to manage a world increasingly reliant on data-driven decision-making. Different types of data require different policy priorities, and data governance is needed all along the data value chain.

Despite its stated goals to be a regional digital economy leader, Malaysia has not yet developed a comprehensive data strategy. Data use is important in driving innovation and growth. Digital rights, such as privacy and anti-discrimination, are important considerations when developing digital policy and data policy. Public interest technologists who prioritise these issues should be included in such policy-making discussions and decisions.

References

Chen, Le, Ruijun Ma, Anikó Hannák, and Christo Wilson. 2018. 'Investigating the Impact of Gender on Rank in Resume Search Engines'. In *Proceedings of the 2018 Chi Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3173574.3174225.

Cory, Nigel, and Luke Dascoli. 2021. 'How Barriers to Cross-Border Data Flows Are Spreading Globally, What They Cost, and How to Address Them'. International Tecnology & Innovation Foundation. https://itif.org/publications/2021/07/19/how-barriers-cross-border-data-flows-are-spreading-globally-what-they-cost.

Dastin, Jeffrey. 2018. 'Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women'. Reuters. October 11, 2018. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.

Gratian, Margaret, Sruthi Bandi, Michel Cukier, Josiah Dykstra, and Amy Ginther. 2018. 'Correlating Human Traits and Cyber Security Behavior Intentions', Computers & Security, 73:345–58. https://doi.org/10.1016/j.cose.2017.11.015.

Khazanah Research Institute. 2021. #NetworkedNation: Navigating Challenges, Realising Opportunities of Digital Transformation. Kuala Lumpur: Khazanah Research Institute.

Personal Data Protection Act 2010. 2010. https://www.pdp.gov.my/jpdpv2/assets/2019/09/Personal-Data-Protection-Act-2010.pdf.

Taylor, Richard D. 2020. '"Data Localization": The Internet in the Balance' 44 (8):15. https://doi.org/10.1016/j.telpol.2020.102003.

UK National Data Strategy. 2020. 'National Data Strategy'. https://www.gov.uk/government/publications/uk-national-data-strategy/national-data-strategy.

Zuboff, Shoshana. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs.