Public-Private Data Collaboration Case Study

1. INTRODUCTION

In Nov 2019, a private-sector organised Datathon brought together six data contributors from the private and public sectors to test and validate two concepts:

a. That public and private data sharing can take place within a trusted governance framework and in accordance with the relevant regulations; and
b. That fusion of public and private datasets can uncover innovative insights, which can in turn drive useful social and potentially commercial outcomes.

To that end, the Datathon successfully fulfilled these outcomes.

While the data sharing was carried out within a very specific context, this case study aims to share the important lessons drawn from working with regulators, public and private sector data contributors, in overcoming challenges that have traditionally stood in the way of data sharing. This case study steps through the journey that participating organisations underwent, using the **Trusted Data Sharing Framework**[1] as a guide for the process.

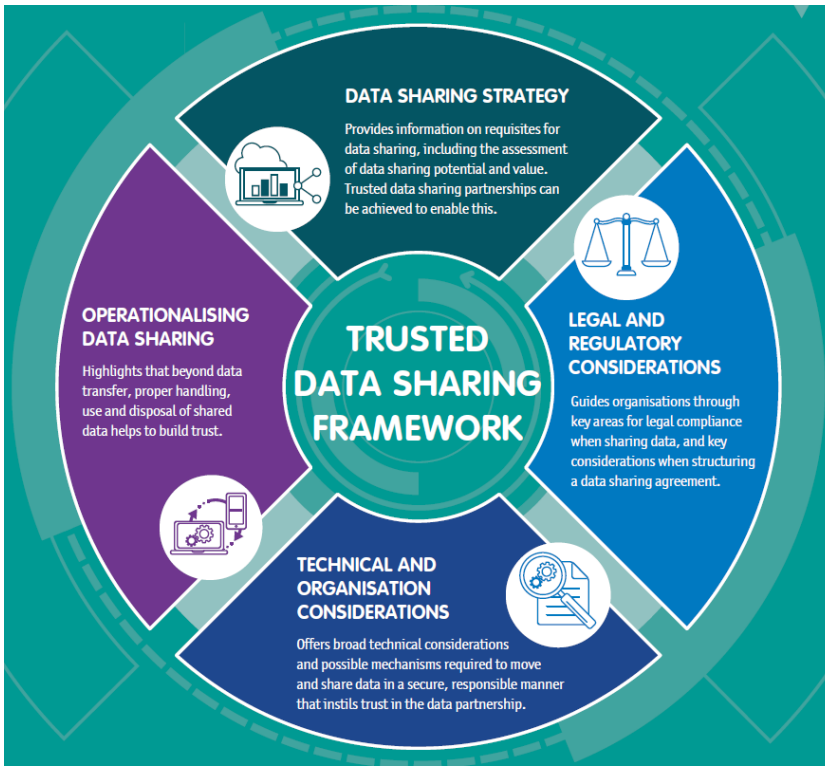

Figure 1: Four Components of the Trusted Data Sharing Framework

[1] IMDA's Trusted Data Sharing Framework can be accessed at https://www.imda.gov.sg/-/media/Imda/Files/Programme/AI-Data-Innovation/Trusted-Data-Sharing-Framework.pdf

The case study also shares:

- What motivated organisations to participate in this data collaboration;
- How the legal concerns were addressed;
- What technical safeguards were put in place; and
- How the data and data products were handled after the event.

2. PURPOSE OF THE DATA COLLABORATION

The participants of this data collaboration believed that the ability to share and harness data would yield comparative and competitive advantages. Fused public and private datasets could be used to uncover innovative insights, which in turn could improve social outcomes for the government, and unlock commercial value for the private sector. At the same time, the data collaboration set out to explore and define the parameters for trust and confidence, such as trusted governance framework and ethical use of data, which would make data sharing possible.

3. DATA SHARING STRATEGY

A few key strategies made the data collaboration possible.

Firstly, the data collaboration focused on use cases with social impact. The proposed cases focusing on health and financial planning for consumers were worthy causes for the data collaboration. At the same time, the potential commercial opportunities that could result were tangential to the data contributors' core businesses, and hence these commercial considerations would not be roadblocks for collaboration.

Secondly, the government and regulator's involvement gave confidence to data contributors to participate. The private sector data contributors were persuaded to contribute data to the Datathon as it was a joint initiative between government agencies, the personal data protection regulator and the private sector. The government led by example, by also contributing individual-level datasets to the Datathon. In addition, data contributors felt assured that the government was the data intermediary to handle shared data, within strict parameters agreed by data contributors.

Thirdly, there was an upfront agreement to destroy the data, fused data and data products generated after the event. This addressed concerns around the equality of the value of contributed data, future access to the fused datasets and the rights to the data products generated through this collaboration.

Lastly, the discussions were conducted across different parties on the different aspects of the project:

- The discussions started first with business decision makers, to identify business outcomes;
- This was followed by discussion amongst data scientists and business units that were the custodians of consumer data, to address the feasibility of the venture;
- Final discussions were handled by technical teams, which focused on addressing the technical details and parameters, e.g. the datasets within and across the participating partners that could be useful for the use cases, the number of records required for the creation of a meaningful fused datasets, the minimum period required, and mechanisms to identify common customers.

4. LEGAL AND COMMERCIAL ARRANGEMENTS

In order to get the collaboration going, all the data contributors signed an agreement that governed critical aspects of the data collaboration, such as specifying the rules of engagement, responsibilities and liabilities of all the parties involved (i.e. data contributors, data service intermediary, data processor and platform provider).

A legal representative from the lead organiser was involved in the discussions with data contributors from the initial stage, to understand the legal requirements and concerns for this collaboration. Early involvement from legal also helped to expedite the drafting of the data collaboration agreement, which addressed most of the business concerns upfront, and provided assurance to the data contributors clearly in the following areas:

i) Scope of data to be contributed and the related rights & liability of various parties, including the data intermediary;
ii) Rights, liability and use of fused datasets and data products;
iii) Scope of practical guidance provided by the regulator;
iv) Obligations for protection of the data (actual and fused) by all parties involved;
v) Governance control implemented by data intermediaries; and
vi) Consequences of termination or expiry of the agreement (e.g. destroy the data upon termination or expiry).

5. TECHNICAL ARRANGEMENTS

To address concerns of data protection and reputation risks for some of the data contributors, synthetic datasets were also created for the purpose of the Datathon. The synthetic datasets simulated the real data provided by the data contributors but retained its statistical properties. All parties agreed that the government would act as the trusted intermediary to collect, fuse and create synthetic datasets.

For the government, as both a data contributor and the trusted intermediary to process and fuse the data, two different teams were set up. This arrangement helped ensure role segregation and maintain the trust from other data contributors. The synthetic datasets were then uploaded and hosted on the commercial cloud-based data sharing and governance platform for access by Datathon participants.

As part of the due diligence process, an external vendor was engaged to perform a re-identification risk assessment, to ensure that the appropriate security protocols had been put in place to effect this method of data-sharing and that the risk of re-identification met the regulatory standards in the Personal Data Protection Act ("PDPA") and the Practical Guidance issued by the Personal Data Protection Commission ("PDPC").

6. DATA REGULATORY SANDBOX & REGULATORY OVERSIGHT

Under the current PDPA, organisations may share personal data where explicit consent has been obtained, personal data has been anonymised or exceptions apply.

The PDPC's guidance was sought on the appropriateness of the proposed data collaboration arrangement through PDPC's **Data Regulatory Sandbox** (see Annex A for more info on the Sandbox). The data contributors sought clarifications on whether there was a need to seek consent from individuals, whose personal data were used in deriving the fused and anonymised / synthetic datasets, and the purposes for which these datasets can be used.

A written practical guidance was provided to address the data contributors' regulatory uncertainties and concerns. The key aspects of the practical guidance provided were:

- Explicit consent was not required for the fusion and anonymisation of data, as long as effective measures were put in place to ensure there was no serious possibility of re-identifying individuals from the datasets and any other information that is (or is likely to be) accessible.

- PDPA did not apply to the collection, use or disclosure of fused and/or synthetic datasets that were anonymised, and such datasets could be used and disclosed for any purposes.

Refer to Annex B for the details on Practical Guidance Sought by Organisations Involved in the Data Collaboration Arrangement.

The involvement of the regulator in the data collaboration process gave assurance to the data contributors that the data collaboration arrangement was compliant with the existing personal data protection law.

7. OUTCOMES AND LESSONS LEARNT

The Datathon brought together real-world data contributors from the public and private sectors, involved the participation of 60 data scientists from than 30 organisations, including universities, start-ups, government agencies and large companies, with support from more than 20 mentors and experts – and uncovered useful ideas on how the data can contribute to social outcomes.

More importantly, it provided useful lessons for future data collaborations:

- Building strong partnerships by bringing together data contributors who understand the need to collaborate and experiment with their data;
- Addressing the interoperability requirements, e.g. data standards, data readiness, common identifiers;
- Developing the data sharing mechanisms, e.g. process for anonymisation, method for fusion, security for transfer, risk assessment; and
- Establishing the legal framework for data sharing, e.g. regulatory clarity around consent and anonymisation, roles and responsibilities of all partners in the data collaboration.

8. SUMMARY

The entire journey took two months: from getting the data contributors to participate, signing the data collaboration agreement, receiving practical guidance from the regulator, extracting data from all the contributors, fusing the datasets, creating synthetic datasets and making the datasets available on a platform. The data collaboration demonstrated how the challenges – i.e. (a) establishing legal framework, (b) resolving regulatory uncertainty with respect to the anonymisation process, and (c) addressing lack of trust via a data service intermediary and its platform – can be addressed in a short time if partners are willing to come together and bring their different skills and expertise to innovatively collaborate and solve the issues.

ANNEX A: DATA REGULATORY SANDBOX

Data Regulatory Sandbox allows businesses and their data partners to explore and pilot innovative use of data in a safe environment, in consultation with IMDA and PDPC. The sandbox reduces uncertainty in compliance to current and planned policies, and limits the exposure of companies and consumers.

There are three stages in the data regulatory sandbox:

- Engagement
- Providing Guidance
- Policy Prototyping

The stages are not necessarily sequential, and dependant on the company's use case and readiness.

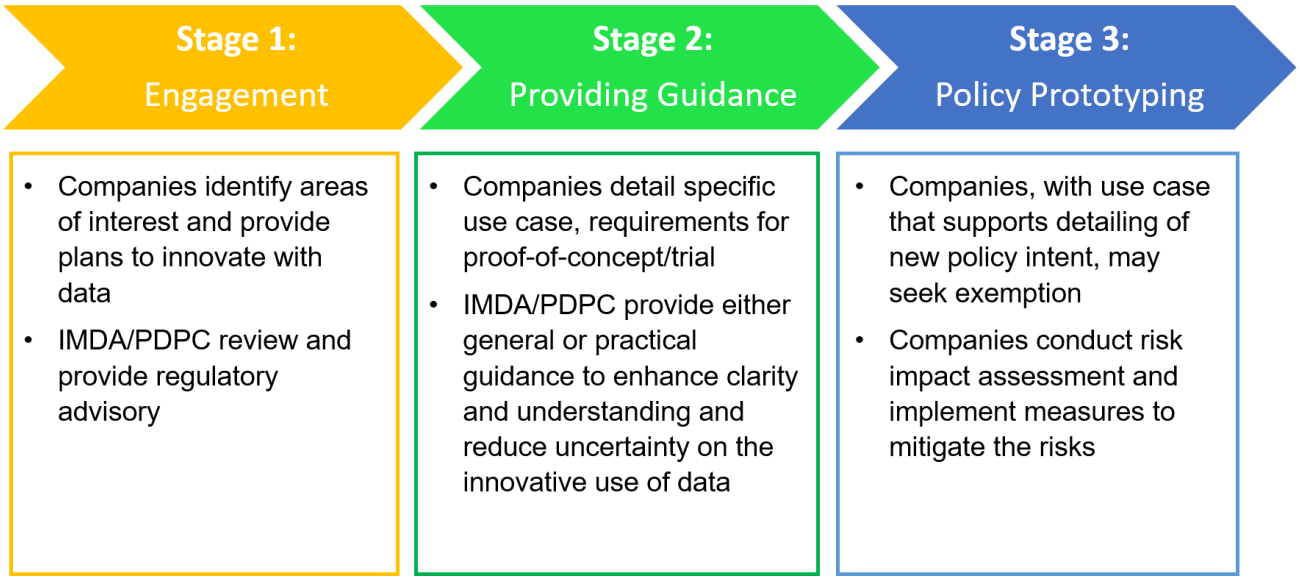

Figure 2: Three Stages of the Data Regulatory Sandbox

ANNEX B: PRACTICAL GUIDANCE SOUGHT BY ORGANISATIONS INVOLVED IN A DATA COLLABORATION ARRANGEMENT

1. Guidance from the Personal Data Protection Commission (the "Commission") was sought in relation to a data collaboration arrangement involving several organisations and a public agency (collectively referred to in this Guidance as "data collaboration partners"). The purpose of the data collaboration is to address social well-being issues through the use of anonymised or synthetic datasets. The anonymised or synthetic datasets will be prepared by a dedicated team from the public agency, using pseudonymised data[2] provided by the data collaboration partners.

2. The data collaboration arrangement will involve the following steps:

   (i) **Defining population of interest:** There will be one team ("Team A") from the public agency that defines the parameters for a population of interest.

   (ii) **Generating pseudonymised data using a common salt:** Team A will then define a common salt, which will be used with an agreed irreversible encryption algorithm to generate a one-way hash.

   (iii) **Sharing of population of interest and common salt with data collaboration partners:** Team A will share (a) the parameters for the defined population of interest; and (b) the common salt with its data collaboration partners.

   (iv) **Extracting the relevant data and generating pseudonymised datasets**: Based on the parameters and the common salt shared by Team A, the data collaboration partners will select a sample dataset from their database to ensure sufficient overlap with the population of interest, and then apply the one-way hash to generate pseudonymised datasets relating to the selected individuals, consisting of hashed identifiers and the corresponding data points that are relevant for the data collaboration.

   (v) **Sharing of pseudonymised datasets to a dedicated team for fusing and anonymising of datasets:** The data collaboration partners, including Team A, will then share the pseudonymised datasets with a separate team from the public agency ("Team B") that is in charge of fusing and anonymising the datasets on behalf of the data collaboration partners. There will be strict

[2] Pseudonymisation refers to the replacement of identifying data with made up values. More details on pseudonymisation can be found in PDPC's Guide to Basic Data Anonymisation Techniques.

separation of roles between Team A and Team B, and Team B will not have access to the common salt that generated the hashed identifiers.

(vi) **Fusing and anonymisation of datasets**: Team B will then fuse the pseudonymised datasets provided by all data collaboration partners. If the fused dataset is not assessed to be anonymised, Team B may further process the fused dataset to ensure it is anonymised. If the fused dataset cannot be anonymised, Team B will generate a synthetic dataset from the fused dataset.

(vii) **Disclosing anonymised and/or synthetic datasets**: The fused dataset and/or synthetic dataset will be disclosed to another organisation for purposes of hosting the datasets for others to access and analyse for a set of pre-defined purposes. The datasets will be hosted in an environment that disallows downloads.

3. The Commission understands that there may not be consent obtained for the collection, use and disclosure of individuals' personal data in a pseudonymised format for the purpose of this data collaboration arrangement.

PDPC's Guidance

4. Specifically, the Commission provided guidance on the following:

(i) Whether consent is required for the data collaboration partners **to disclose** pseudonymised data to Team B to fuse the data and to anonymise the fused datasets (Data Extraction and Pseudonymisation stage);

(ii) Whether consent is required **for processing** of pseudonymised data to fuse and create anonymised datasets (Data Fusion and Data Anonymisation stage); and

(iii) Purposes for which the fused dataset and/or synthetic dataset **may be used and disclosed** (Data Distribution stage).

5. For the avoidance of doubt, the guidance set out in this document has been scoped to address the situation as described in paragraphs 1 to 3 above, based on the information provided.

6. The Commission's guidance relates to the application of the Personal Data Protection Act 2012 ("PDPA"), and is not applicable to the public agency[3] involved in the data collaboration arrangement.

[3] Refer to Section 4(1)(c) of the PDPA; further, data-related activities of public agencies are governed under the PSGA.

(a) Whether consent is required for data collaboration partners to disclose pseudonymised data to Team B to fuse the data and to anonymise the fused datasets (Data Extraction and Pseudonymisation stage);

7. The Commission notes that under the proposed data collaboration arrangement, Team B is fusing and anonymising the pseudonymised data on behalf and for the purposes of the data collaboration partners. Given so, data collaboration partners are responsible for ensuring that there are appropriate safeguards (e.g. legal and procedural controls) in place to ensure the fusing and anonymisation of data done by Team B are scoped to their purposes and individuals cannot be re-identified from the anonymised and/or synthetic datasets, in accordance with section (c) below.

8. Where Team B is from a public agency, it is excluded from the application of the Data Protection Provisions of the PDPA. Nonetheless, other data collaboration partners' disclosure of personal data to Team B is subject to the PDPA. Under the PDPA, consent is required for the organisations to collect, use and disclose personal data for a purpose (unless an exception applies). However, **consent is not required to process and convert personal data into anonymised data**. The Commission is of the view that **the data collaboration partners' disclosure of pseudonymised data to Team B is for the purposes of fusing the data and anonymising the fused dataset, for which separate consent is not required.** As consent is not required for the fusing and anonymisation of data, consent is consequently not required for the data collaboration partners' disclosure of pseudonymised data to Team B to perform the fusing and anonymisation.

9. **The analysis in the preceding paragraph applies in the situation where Team B is not a public agency and is a data intermediary[4] under the PDPA.** In this case, it will be subject to the Protection[5] and Retention Limitation[6] Obligations under the PDPA. In the event the data intermediary uses or discloses personal data in a manner that goes beyond the processing required by the data collaboration partners (e.g., using or disclosing data obtained from the data collaboration partners for other purposes), it will be required to comply with all Data Protection Provisions under the PDPA. Among other requirements, it must ensure

[4] A "data intermediary" is defined in the PDPA as an organisation that processes personal data on behalf of and for the purposes of another organisation. The arrangement where an organisation is to act as a data intermediary of another organisation should be set out clearly in a contract that is evidenced or made in writing, including the responsibilities and liabilities of each organisation in relation to the processing of the personal data.

[5] Section 24 of the PDPA.
[6] Section 25 of the PDPA.

that consent has been obtained to use or disclose the personal data for these other purposes.

(b) Whether consent is required for processing of pseudonymised data to fuse and create anonymised datasets (Data Fusion and Anonymisation Stages)

10. The Commission notes that the proposed data collaboration arrangement will incorporate the following measures as safeguards during the course of the collaboration:

    (i) Use of hashed identifiers based on a common salt;

    (ii) Separation of Team A (which defines the common salt) and Team B (which fuses the data and anonymises the fused dataset). This ensures that Team A will not gain access to pseudonymised datasets shared by the other data collaboration partners, and Team B will not have access to the common salt used to generate pseudonymised datasets; and

    (iii) Contractual agreement between data collaboration partners clearly specifying Team B's obligations and responsibilities in respect of the fusing and anonymisation of datasets, on behalf of and for the purposes of the data collaboration partners.

11. The Data Protection Provisions under the PDPA do not apply to anonymised data. "Anonymisation" refers to the process of converting personal data into data that cannot be used to identify any particular individual, and can be reversible or irreversible[7].

12. The Commission notes that at the Data Fusion and Data Anonymisation stages, the outcome of Team B's processing is to create fused and/or synthetic datasets that are anonymised. Based on the position stated in paragraph 8, **data collaboration partners do not need to obtain consent to fuse the data and to anonymise or create synthetic dataset from the fused dataset.**

(c) Purposes for which the fused dataset and/or synthetic dataset may be used and disclosed (Data Distribution stage)

13. The Data Protection Provisions under the PDPA do not apply to the collection, use or disclosure of fused and/or synthetic datasets that are anonymised, and such datasets can be used and disclosed for any purposes. This would include **the use and disclosure of fused and/or synthetic datasets that are anonymised before the Data Distribution Stage, as well as for research, data mining, data**

[7] Refer to PDPC's Advisory Guidelines on the PDPA for Selected Topics on Anonymisation.

analytics, development of commercial products and services, developing Artificial Intelligence machine learning models, etc.

14. In general, data collaboration partners should assess the risks of re-identification of individuals from the resultant datasets. They should put in place effective measures to ensure there is no serious possibility of re-identifying individuals from the datasets and any other information that is (or is likely to be) accessible. If there is a possibility that an individual can be re-identified from the fused and/or synthetic dataset, the Commission will not consider that dataset to be anonymised, and consent would be required for the collection, use or disclosure of such datasets. Please refer to PDPC's Advisory Guidelines on the PDPA for Selected Topics (Chapter 3 on Anonymisation) for further information on assessing and managing the risks of re-identification of anonymised data.

END OF DOCUMENT