DATA71011 Understanding Data and their Environment 2022-23

Individual Essay

Disclosure Data Risk and Mitigation

A case study of the Troubled Families Data situation

Yanfei Shan: 11118569

1. Introduction

Disclosing or sharing data to another party can have dual implications. On one hand, it facilitates collaboration between different organizations, drives innovation and research and improves decision-making. On the other hand, it also brings risks including privacy violations, unauthorized access, misuse of information, and theft of sensitive information. The consequences of such risks can range from minor financial losses to significant harm to individuals, organizations, and society. Thus, it is becoming more and more critical for organizations to weigh the possible risks before they disclose data. This article will use troubled family's data, which is highly sensitive, as an example through the Troubled Families Programme(TFP) which allows researchers access to the data, to analyze the risks involved and provide appropriate recommendations.

2. Background and Methodology

The Troubled Families Programme(TFP), which was launched in England in 2011 and ended in 2019, aimed to provide targeted support to families facing multiple and complex problems such as poverty, unemployment, poor educational achievement, and poor health(especially mental health). Many years of operation have given the TFP rich data, which was believed giving access to researchers who can gain greater insight and inform future policy and practice (e.g. Building Better Futures programme). However, getting more people involved in the project will inevitably lead to greater data risk.

Our risk analysis was based on filling in the Anonymisation Decision-Making Framework (ADF)(Elliot M, 2016), where Content 1-4 helped us to catch the changes in the data environment and the possible risks associated with this changing, which are summarised below as Scenario 1. In Content 6, we found that the TFP situation fell in the red zone, i.e. 'Essential ', indicating the importance of risk assessment and process controlling. Content 7 helps us with risk at the data level, which is summarised in Scenario 2. The specific ADF tables are attached in the appendix.

2.1 Assumptions

Before risk evaluation, we make the following assumptions.

- We assume that the current time point is 2022.
- We asscume the data we want to disclose to researchers is the data for each year of the WHOLE programme, i.e. 2011-2019, implying a DATA YEAR between 3 and 12 years.
- We assume that the intruder acts rationally and does not randomly engage in sabotage
- We assume that intruder's motivation is only two factors, financial and political.

3. Risk Analysis

Data risks primarily exist in two major dimensions: environmental, and data itself. In what follows, we present these two scenarios respectively.

3.1 Scenario One (Environment Risk)

Let's say there is a super-rich consortium who want to prove that the TFP is a wrong government

project to support his election. Hence, they hired hackers to access the TFP data directly and used it to generate statements in their favour. The hackers wanted to maximize their profits and get the most complete data with the least amount of effort, then they conducted an environmental analysis of the TFP data.

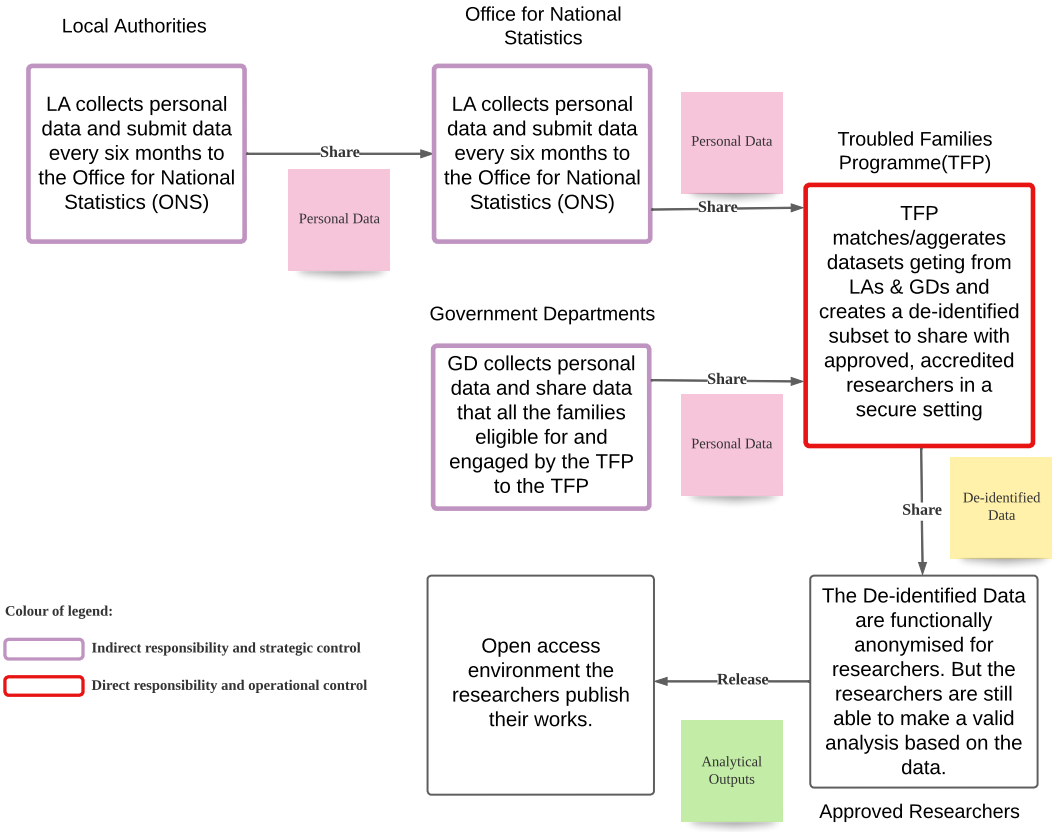


Figure 1. Data flow in different environments

TFP data is not fixed but flows through different environments. Government departments and local authorities[1][2][3][4]only provide a portion of the data, having indirect responsibility and strategic control. Access to the government departments and local authorities' datasets is strictly controlled, and only authorized individuals and organizations are permitted to use the data for approved purposes. TFP combine the information together, having direct responsibility and operational control on what data to choose. This means that before the TFP shares data with researchers, the data is closed and controlled.

As rational hackers, first they would prefer to access a closed database rather than an open data set. This is because closed databases typically contain more sensitive and personal information, such as names, addresses, and other identifying details, that can be used to reidentify individuals. In contrast, open data sets are often stripped of sensitive information or otherwise de-identified to minimize the

[1] Data Origin: combines administrative data from multiple government departments and local authorities. (On around 60,000 families data recorded)

[2] The Police National Computer (PNC) held by Ministry of Justice

[3] The National Pupil Database (NPD) held by Department for Education and the Work and Pensions Longitudinal Study (WPLS) held by Department for Work and Pensions

[4] Family Progress Data (FPD) by local authorities directly to DCLG

risk of reidentification. Thus, they will target the TFP environment and researchers as the first tier of attack.

As for the environments before TFP, although they were also closed data sets, they were dispersed and attacking each one individually would cost a lot. On the contrary, the data of TFP and researchers is already aggregated and specific to this programme. Furthermore, data owned by researchers has higher possibility to be attacked. TFP data was stored and managed by the Department for Communities and Local Government (DCLG), with more safety infrastructure to prevent intruder's attack. Once they share the data to researchers, these safety infrastructures cannot guarantee that the researchers' data will not be leaked. As more researchers involved in this programme, if one of their databases is vulnerable, hackers will be able to get the data they want. (e.g., Zimmer 2010)

3.2 Scenario Two (Data Risk)

In Scenario 1, we envisaged the possible risks posed by the environment, in this section, we analyse the risks of the data itself. We imagine that the intruder is a criminal gang that sells information. They sell illegally obtained personal data to insurance companies or fraud groups for profit. Their technical means are limited, but they have plenty of time and spend a lot of time analyzing publicly available data sets and reports. The main reason they chose to attack TFP is that this contains a lot of information about poor families, which can be used extensively for financial fraud, e.g., health insurance fraud. Health insurance fraudsters have been known to target poor families by submitting false claims for medical treatments or procedures that were never performed.

3.2.1 High risk direct identifier

Through the report and metadata for potential TFP dataset released by TFP, we have marked the following sensitive attributes. Among them, Name is a direct identifier, although there are people with the same name, but with other information, we can easily identify a person.

Table 1. Risk attributes

Variable Type	Variable Meaning
Direct	Local authority name
Indirect	The gender of the child
Indirect	Child's year of birth
Possibly	The number of people in that child's family at the time of their birth
Possibly	The child's age when their family member had their first mental health and/or drug or alcohol episode
Possibly	If lone parent families
Possibly	If family with an individual with any mental health issue
Possibly	If Family with an individual dependent on drugs or alcohol
Possibly	If families who have been involved in a domestic abuse incident

3.2.2 Longitudinal data

In Table 1, the green attributes indicate potential identifiers. Some belongs to the longitudinal attribute, which changing over time. For example, by concatenating the data with the previous year, intruders can locate information on the new birth family and, combined with information on births in the local ward, may be able to obtain information on a local trouble family's information.

3.2.3 Hierachical data

Another class belongs to hierarchical attribute. This type of data may appear in multiple data sets, and if an intruder merges two tables with similar variables, they can find the overlapping portion. For example, if a troubled family has a higher risk of mental illness and drug abusing, then linking the TFP to the local hospital discharge record will allow for reidentification. (Ohm, 2010)

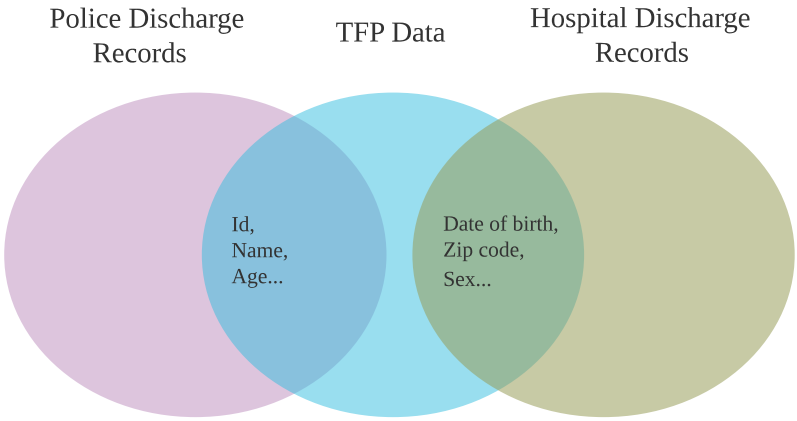


Figure 2. Hypothetical example of reidentification attack of the Troubled Families Programme

3.3 Risk Mitigation

From the discussion of the two scenarios, we conclude that in the case of TFP situation, the main threats come from the researchers, and data that is chosen to be disclosed. To mitigate risks, some steps can be taken include:

- Cryptography and environmental security: Hire network security experts, upgrade database security systems; Secondary passwords permissions for researchers. They are required passwords to access the database, and another certificate to open the data. (Reiter and Kinney 2011)
- Modification to Data: De-identify the data when we share it with the researchers, removing the direct identifier, differencing the data or applying k- anonymity (Sweeney 2002b), and turning a specific attribute, such as age, into an interval.
- Publicity: keep a low profile when bringing in researchers. Low-profile research is less likely to be noticed by these opponents and may be a less attractive target.
- Ethical Appeal: Each participant and researcher should read and sign the statement defining how he or she would use the data responsibly.
- Regular security assessments: Conducting regular security assessments and penetration testing can help identify and address vulnerabilities in the database.

4. Conclusion

This case study analysed the possible risks among disclosing the TFP data to researchers in

environmental and data itself aspects. When deciding share data to another party, we should take the risks it may occur. There are many approaches in different areas can be used to mitigate the risks. We hope this case study showing the possible that safe and responsible data access will be ensured through collaboration from different sectors to improve the technical, legal, ethical, and social frameworks.

Reference

Mackey E, Elliot M, O' Hara K. The anonymisation decision-making framework[M]. UKAN publications, 2016.

Ohm, Paul.2010. "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization." UCLA Law Review 57(6):1701.

Kinney, Satkartar K., Jerome P. Reiter, Arnold P. Reznek, Javier Miranda, Ron S. Jarmin, and John M. Abowd. 2011. "Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database." International Statistical Review 79(3):362–84.

Sweeney, Latanya. 2002b. "k-Anonymity: A Model for Protecting Privacy." International Journal of Uncertainty, Fuzziness and Knowledge-based Systems 10(05):557–70.

Zimmer, Michael. 2010. " 'But the Data Is Already Public' : On the Ethics of Research in Facebook." Ethics and Information Technology 12(4):313–25.

ADF on the Troubled Family Programme Data

ADF COMPONENTS	
Capture the presenting problem **Issues to think about to capture the presenting problem:** – Who are you? What role do you play? – What is the use case? – **Why?** – **Who?** – **What?** – Is the environments fixed? – Is the data fixed? –	I am a postgrad student from University of Manchester. My role is to assess the risk of allowing more researchers to join in Troubled Family Programme (TFP) analysing data. And give appropriate mitigations according to this situation. The purpose of allowing researchers to access TFP's data is to gain a deeper and more professional understanding of the complex issues faced by troubled families and uncover patterns and relationships that may not be immediately apparent from the data collected by government departments and local authorities. At the same time, the involvement of experts can propose more efficient, cost-saving, and effective methods to help troubled families. The researchers who work on the TFP could have been from a variety of organizations, including government departments, universities, think-tanks, research institutions, and non-profit organizations. They might have been based in the United Kingdom or from other countries. Meanwhile, they might come from a range of academic disciplines, including social policy, sociology, economics, public health, and education, among others. Researchers may apply multiple methods to analyse to the dataset and conclude from findings which may help policy decisions. The environment is not fixed. The data flowed in different environments, with some local authorities collecting and storing the data, sharing them with the Ministry of Housing, Communities and Local Government (MHCLG), which in turn shared the data with researchers. Researchers then produce statistics and reports which are returned to the MHCLG, which then publishes them to the general public. The data is not fixed, as the situation of the troubled families in England is changing and some of the values of the variables in the data set have changed according to the implementation of the TFP, for example, more people have gotten jobs. Therefore, this data is a changing data set and researchers need to do new analysis and generate new reports based on the new data set each year.

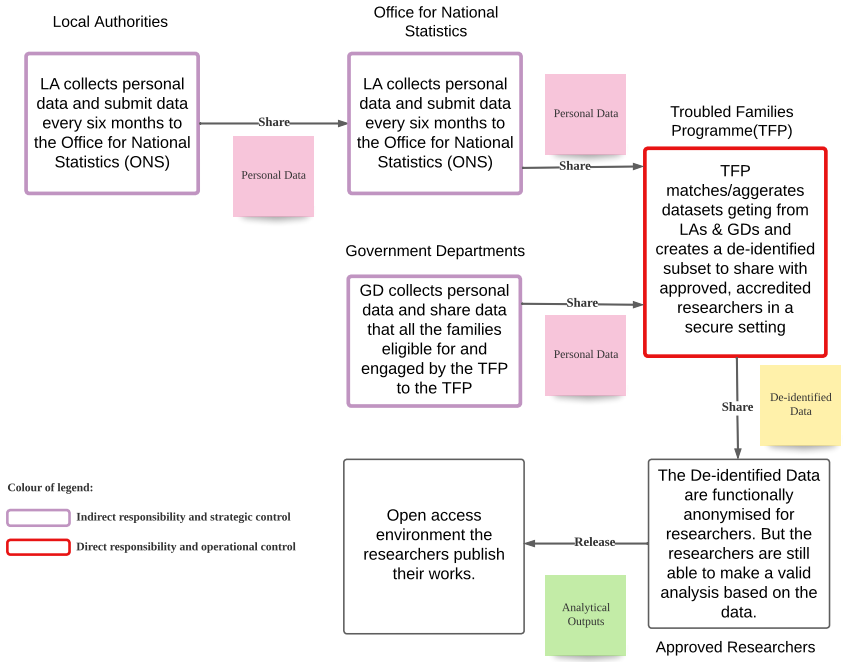
Sketch the data flow **Issues to think about to sketching the data situation:** – Where has the data come from – Where is the data going to – What does the data flow look like – Over what environments do you have control and/or responsibility? – What parts of the data situation are relevant? *Use a separate page to sketch flow if there is not enough room here.*	Data Origin: combines administrative data from multiple government departments and local authorities. (On around 60,000 families data recorded) Several sources: ✧ The Police National Computer (PNC) held by Ministry of Justice ✧ The National Pupil Database (NPD) held by Department for Education and the Work and Pensions Longitudinal Study (WPLS) held by Department for Work and Pensions ✧ Family Progress Data (FPD) by local authorities directly to DCLG Data is going to researchers for deep analyses, and researchers are going to publish the outcomes they get from the data. Data Flow: Local authorities collect identified information from data subjects and disclose the collated data to the relevant government departments, which then aggregate the data and share it with TFP. Then, TFP aggregates data from multiple government departments & local authorities, de-identified data and filters useful information to create a new dataset. Then this new dataset is shared with researchers. Government departments and local authorities have indirect responsibility and strategic control. TFP has Direct responsibility and operational control. Access to the government departments and local authorities' datasets is strictly controlled, and only authorized individuals and organizations are permitted to use the data for approved purposes. This means that before the TFP de-identifies the dataset and shares it with researchers, the data is controlled. However, after the researchers obtain the data, the TFP has limited say on what researchers do.

Map the properties of the data environment(s)

Issues to think about to sketching the data situation:

- Which environments are relevant?
- Have you captured all the features?
- Have you properly understood the boundaries of the DEs?

Relevant environments:



- Other Data: Access to other data is strictly controlled, and only authorized individuals and organizations are permitted to use the data for approved purposes.
- Agents: Specified group
- Governance: Access controls, output checking, purpose limitation
- Infrastructure: Software

Describe and Map the data **Issues to think about to establish a risk profile:** – Data structure (text or numerical) – Data type (microdata or aggregate) – Data population type (sample or whole) – Variable type (identifiers, keys, special category data) – Dataset property type (relationships in the data structure, quality, size of dataset, cross sectional or time series) – Topic area type (sensitive or non-sensitive) **Complete the Data features form (see handouts).**	Data Structure: Most of variables are numerical, several variables are text, like gender, name. Data type: a mix of microdata and aggregate data Data population type: sub-population (The families eligible for and engaged by the programme) Variable type see Appendix 3. Dataset property type: ✧ Data Quality: The data comes from different local authorities and government departments, presenting a high degree of confidence. However different departments measure and match the data differently, and TFP matches the data through post-adjustment, so the data does not have a particularly high level of completeness and may has a small level of error when doing the match. ✧ Age of Data: The data used to generate National evaluation of the Troubled Families Programme 2015 – 2020 is from 2017. However the data is updated every 6 months. ✧ Hierarchical data: Age, Gender ✧ Longitudinal data: Family Size Topic Area Type: Sensitive

The Data situation Evaluation For the Troubled Family Progamme

A. Agreement Sensitivity – For the purpose of this exercise, 'agreement' is defined as an active indication of a preference, which maybe expressed explicitly e.g. given verbally or in script such a signature or symbol or implicitly by participation in an activity. The concept captures also the notion of an agent's awareness and opportunity to express a preference.	
1. Are the data subjects aware that their data have been collected in the first place?	Yes
2. Have the data subjects agreed (explicitly or implicitly) to the collection of their data?	Yes
3. Were the data subjects completely free to agree to the collection of their data (or have they agreed to collection because they want something (a good or service) for which are required to hand over some data before they can obtain it)?	Yes
4. Are the data subjects aware of the original use of their data?	No
5. Have the data subjects agreed (explicitly or implicitly) to the original use of their data?	Yes
6. Have the data subjects agreed in general to the sharing of a functionally anonymised version of their data?	Yes
7. Are the data subjects aware of the specific organisations that you are sharing a functionally anonymised version of their data with?	No
8. Have they agreed to your sharing their data with those organisations?	Yes
9. Are the data subjects aware of the particular use to which their functionally anonymised data are being put?	No
10. Have they agreed to those uses?	Yes
A: Count the Number of No's	3

B. **Expectation Sensitivity**	Yes, No or N/A
1. Does your organisation have a relationship with the data subjects such that a reasonable data subject would expect you to have access to their data?	Yes
2. Does the receiving organisation have a relationship with the data subjects such that a reasonable data subject would expect them to have access to their data?	Yes
3. Is the receiving organisation a government or commercial entity?	Yes
4. Is your organisation's area of work one where trust is operationally important (e.g. health or education)?	Yes
5. Will you receive financial or commercial benefit from the data share?	No
6. Is there an actual or likely perceived imbalance of benefit arising from the proposed share or release? e.g. is the data controller benefiting but the data subjects not?	No
B. Add the Number of Yes's to questions 3-6 and the number of no's to questions 1 and 2 and then multiply by 2.	2

C. Data sensitivity	Yes/No
1. Are some of the variables sensitive?	Yes
2. Are the data about a vulnerable population?	Yes
3. Are the data about a sensitive topic?	Yes
4. Is the use of the data likely to be considered sensitive?	Yes
5. Do you have reason to believe that the intended use of the data might lead to discrimination against the data subjects or a group of which they are members?	Yes
C. Number of Yes's multiplied by 2	10

D. Desensitising Factors	
1. Will there be some public benefit arising from the downstream use of the data? (Yes = -3, No =0)	No
2. Have you carried consultations with groups of stakeholders (particularly the general public and/or data subjects)? (Yes = -3, No =0)	Don't Know
3. Have you carried consultations with groups of stakeholders (particularly the general public and/or data subjects) and implemented the recommendations arising there from? (Yes = -10, No =+3)	Don't Know
4. Does your communication plan engender trustworthiness through transparency (sufficient to offset adverse responses in the expectation sensitivity section)? (Yes = -5, No =0).	Don't Know
Total Data situation Sensitivity A+B+C+D	3

If your total score is 4 or less, data situation sensitivity is *Low*. If it is between 5 and 10 it is *Moderate* and above 10 it is *High*. Use this classification in combination with the summary likelihood onto which we will now move.

Summary Risk	points
1. Are the data of high quality?	2
a. Yes, the data are clean and contain no or minimal errors and no or minimal missing data. (2 points)	☑
b. The data contained errors but have been cleaned. (1 point)	
c. The data contains some errors and or missing data. (1 point)	
d. The data are dirty - they contain many errors and missing data issues. (0 points)	
2. How old are the data?	3
a. Less than 1 year. (5 points)	
b. 1-5 years. (4 points)	
c. 5-10 years. (3 points)	☑
d. 10-20 years. (2 points)	
e. More than 20 years old. (0 points)	
3. Do the data constitute a whole population or a sample?	4
a. Population. (5 points)	
b. Sub-Population (4 points)	☑
b. Sample. (0 points)	
4. How many variables are there that fall within the standard key variable sets?	1
a. 0. (0 points)	
b. 1-4. (1 point)	☑
c. 5-9. (4 points)	
d. 10+. (5 points)	
5. Which of the following best describes the data?	1
a. A single aggregate output. (0 points)	
b. A set of aggregate outputs that do not overlap. (1 point)	☑
c. A set of aggregate outputs that do overlap. (4 points)	
d. Flat microdata. (4 points)	
e. Hierarchical but not longitudinal[1] microdata. (7 points)	
f. Longitudinal but not hierarchical microdata. (7 points)	

[1] Note that avoid double counting if all the temporal information in a datasets is captured in question 6 then do not count as longitudinal here.

g. Hierarchical and longitudinal microdata. (10 points)	
6. Do the data include any data types that present particular reidentifiability challenges (e.g. genomics data, photographs, significant text narratives, timestamped location data or other timestamped sequences)?	0
No/Yes (0/10 points)	No
7. Now considering the details of the focal environment, which of the following best describes that environment?	-20
a. It is a remote analysis server where users may submit code for analyses but are not able to directly access the data. Code and output are checked before the outputs are released to the user. (-25 points)	
b. It is a secure facility with on-site access with limited personnel being able to access the data that is housed within the data controller's infrastructure (-25 points)	
c. It is a secure facility owned by the user. (-20 points)	
d. It is a remote access server with controls on the who and how of access. Users will be able to interact with the data but do not have a copy themselves (so are prevented from linking to other datasets). How users access and work with the data is pre-specified. (-20 points)	☑
e. It is a point-to-point data share based on a bespoke data sharing agreement(s) with purpose limitations, data minimisation, and specific named users. Some auditing for compliance is in place. (-5 points)	
f. It is a licensing environment. Users sign a license agreement to access the data and then are able to download them. (-2 points). Restrictions and policing of secondary use are limited.	
g. The environment is open or quasi-open (with minimal sign up conditions). (0 points)	
8. Are there data in - or which could be moved into - the focal environment that could be used to re-identify any data subjects in the data?	5
Yes/No/don't know (10/0/10)	Yes
Points Total	6