FIT5205 ASSIGNMENT 2

PRIVACY IMPACT ASSESSMENT

SOVAN SINHA ROY

30201152

***TUTOR*: MOZHDEH DEHGHANI**

1) BRIEF DESCRIPTION OF THE PROJECT

Priority Investment Approach **(*PIA*)** to Welfare is an initiative by the *Australian Government's Department of Social Services* **(*DSS*)** to better comprehend and help those who are vulnerable and disadvantaged (Services, 2019). The ***PIA*** uses *actuarial (data) analysis* to identify groups at risk of long-term welfare dependence and by virtue of insights accrued from the data analysed find innovative ways to assist more Australians live *independently* thus reducing their need for welfare schemes (Services, 2019). Furthermore, the initiative relies on the *actuarial analysis* of *social security* to address the risk of intergenerational welfare dependence and subsequently reduce long-term *social security costs* (Reddel, 2018).

So, the ***PIA*** consist of a dataset which is a de-identified dataset containing *administrative data* extracted from *Department of Human Services' (DHS) Enterprise Data Warehouse* for the purpose of documenting and recording various service delivery activities to keep track on the expenditure of government revenue in the shape of payment & welfare schemes (Reddel, 2018). Interestingly, in order to permit researchers to *explore aspects of the data* and also to let them *design, develop, & test analytical models*, as well as to render a *teaching tool for academics & concerned community* on *complex social welfare data*, a *Synthetic Dataset* containing several quarterly snapshots has been further produced & published via *data analytics* that is available on *Australian Data Archive's Dataverse* offering informative data with restricted access and shielded privacy of the individuals whose data is somewhat represented (post modifications) (Services & Data61, 2017).

2) CONSIDERATION OF WHETHER THE PROJECT INVOLVES THE COLLECTION, STORAGE, USE OR DISCLOSURE OF PERSONAL INFORMATION OR NOT WITH BRIEF DETAILS

Yes, the authentic ***PIA*** containing *administrative data* does contain or involve the collection of *personal information,* various *geographical & demographic* details are stored and subsequently utilized for several purposes which will be discussed later. However, only limited restricted access to the ***PIA*** data is provided via a *Synthetic Dataset* version for the perusal of people like *researchers* and *academics*, etc (Spokes, 2018). The *Synthetic Dataset* has been meticulously crafted applying *privacy-preserving algorithm* on the original ***PIA*** data which has been reviewed and certified by *independent assessors* (Services & Data61, 2017) and hence a person's true identity is never disclosed, it is inherently altered & modified such that the overall group data very closely resembles that of the original dataset (Spokes, 2018). Further delineation is enumerated as under:

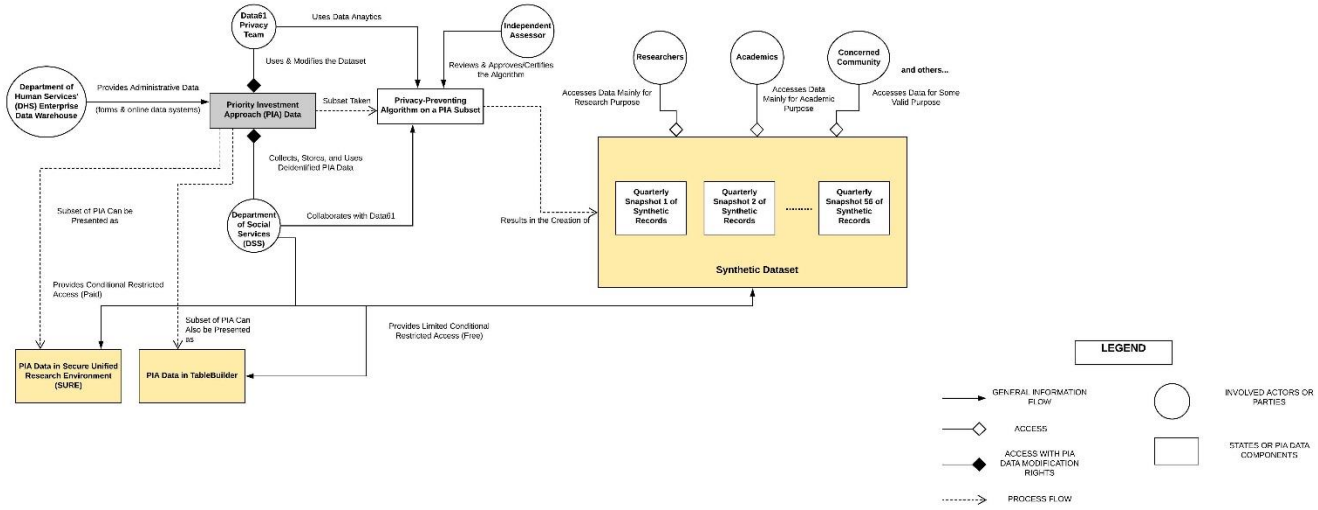
A) So, a myriad of variables involving a person's basic *demographic* and *geographical* and/or *longitudinal information* like *Address, Date-of-Birth, Country of Birth & Residence, Gender, Marital Status, Number of Children,* etc as well information on the *Type Of Accommodation, Primary Medical Conditions, Education, and Income & Income Support,* etc are collected and thereafter stored in a de-identified manner in the original ***PIA*** dataset (Services, 2019). However, in the *Synthetic Dataset,* only limited information (with 31 variables to be precise) is furnished with modified sensitive information pertaining to *Personal ID Number, date-of-birth*, etc. This is done in order to safeguard the privacy of the people whose information was collected (Services & Data61, 2017). The digitally published quarterly records in the *Synthetic Dataset* on *Dataverse* are not readily available for everyone, it is only accessible to restricted legitimate users free-of-charge who have a genuine purpose behind accessing the data. Moreover, no *names* are recorded or disclosed in the *Synthetic Dataset* records and the *longitudinal* aspect is also truncated in it.

B) The main objectives behind the collection & usage of all the essential de-identified information from people are *recording & ascertaining eligibility for benefits, and facilitating service delivery activities and payments* (Reddel, 2018). Additionally, the prime reason for providing access to the thereafter formulated *Synthetic Dataset* is for the ***potential researchers, academics, and concerned community, etc*** *to gain a decent comprehension & insight into the prevalent and upcoming social policy schemes* (Povey et al., 2018).
The *administrative data* were collected from the *Department of Human Services' (DHS)* ***forms & online data systems*** under the purview of the *Department of Social Services* **(*DSS*)** of the *Australian Government* (Services, 2019). 56 quarterly records in the *Synthetic Dataset* were later created by a reliable data innovation group ***Data61*** in collaboration with the ***DSS*** under its supervision (Services & Data61, 2017).
Highly sensitive personal information like *Personal ID Number* et al were collected but since it was stored in a de-identified manner the threats to privacy and data security were substantially alleviated. Furthermore, as no ulterior motives can be ascertained for the data usage by the competent authority; all the *personal, demographic, geographical, and longitudinal information* can be deemed fairly secured under their tutelage. Moreover, to uphold the privacy of the people, *privacy-preserving algorithms* were applied to a subset of original ***PIA*** data using *data-analytics* to concoct the *Synthetic Dataset* which was duly reviewed and approved by *independent assessors* (Services, 2019). However, amidst all this the people whose data had been *extracted, used* and subsequently *disclosed* after amendments are never *notified* nor are they contacted for any sort of *consent*!

C) The key stakeholders of this initiative are identified as follows:
The onus is on the *Department of Social Services* ***(DSS)*** who has invariably undertaken this initiative and supervises it, they are the key stakeholders and the *Australian Government* is also explicitly involved with this initiative. The Australian Data Privacy group ***Data61*** who worked in collaboration with the ***DSS*** while developing the *privacy-preserving algorithm* also has a major interest and influence in the ***PIA***. The *National Innovation Science Agenda* (NISA) that co-funded the project has a massive interest in it too. Moreover, the *researchers, local community & academics,* etc are also imperative stakeholders here in this initiative as they can access and are actually accessing the available data for their *perusal, learning*, and for *designing and testing analytical models* (Services & Data61, 2017). Interestingly, as the data collected from the 5 million people are not presented "*as it is"* nor is it published with the full original information of them affecting them in any significant manner, they shouldn't be considered a stakeholder here as the *Synthetic Dataset* cannot be used to identify any real person.

3) COMPREHENSIVE INFORMATION FLOW

Please **Zoom** it, refer to the last page of the document, or click this link to get a better and clearer picture: https://www.lucidchart.com/invitations/accept/75583106-6fa5-4cd4-8922-eff004bcbcc7



So, to put it in a nutshell, the *Information Flow Diagram* above illustrates how the administrative data is initially extracted from the DHS Enterprise Data Warehouse for the PIA data under the supervision of DSS that ensures only de-identified data is stored and used. Thereafter, the Data Privacy group *Data61* in collaboration with the DSS develops *privacy-preserving algorithm* serving as the catalyst for Synthetic Dataset. An independent assessor reviews and approves the algorithm and eventually Synthetic Dataset is created and published for people like *Researchers, Academics, Concerned Community,* etc who can then access this data digitally free of charge upon request from the *Dataverse* platform of the Australian Data Archive (ADA) for numerous purposes. The different phases of the dataset are also manifested in the diagram connected via a *process flow* representation. Other forms of PIA dataset like *TableBuilder* version & *Secure Unified Research Environment* (SURE) version are also available containing more information about the PIA data but these are beyond the scope for this endeavour so we will not get into them.

4) PRIVACY IMPACT ANALYSIS AND COMPLIANCE CHECK *(Based on available Synthetic Dataset)*

The following *Privacy Impact Assessment* has been undertaken on the *Synthetic Dataset* of the overarching ***PIA*** data taking the **Australian Privacy Principle (APP)** guidelines of the *Office of Australian Information* (OAIC, 2014) into consideration. Furthermore, all the issues and/or reasons raised and discussed briefly in the table below are mostly taken from the *Synthetic Data User Guide* available as part of the *Synthetic data* material (Services & Data61, 2017).

APP	COMPLIANCE STATUS	ISSUES/REASONS	RECOMMENDATIONS
APP 1: Openness and Transparency	**Requires Further Action or Review**	No clarity on the management of the personal information (internally) of the *Synthetic Dataset* from the *privacy-*	More openness and details about the internal management of information are

		preserving algorithm is made distinctly available. A clearly expressed & up-to-date *APP Privacy Policy* is also not available on their *Website, User Guide, or the Data Dictionary* regarding the *Synthetic Dataset* for hindsight on conformance.	required in this regard and are highly recommended. *APP Privacy Policy* adherence details should be made readily accessible on the website or webpage if not the user guide to provide a better picture on the transparency regarding the management of personal information. From the perspective of *Ethical Design Principles* (Beard & Longstaff, 2018), I reckon the deliberations on **Non-Instrumentalism** principle is apt here as ***DSS*** & ***Data61*** team should have done better not to deploy people's information as a tool to achieve their goals and most importantly in that process reduce them to the status of mere 'things' snatching their dignity.
APP 2: Anonymity and Pseudonymity	**Compliant**	Albeit the individuals are not involved explicitly yet all necessary measures have been taken by the ***DSS*** and ***Data61*** team in the *Synthetic Dataset* to ensure that people are *anonymous*, and no one can be identified by others based on scrutinization of the records.	No recommendations
APP 3: Collection of solicited personal information	**Not Applicable**	No sensitive and/or personal information is presented "as it is" seeking consent from the people in the *Synthetic Dataset*. The information extracted was already de-identified and the *Synthetic*	No recommendations

		Dataset underwent further transformation of record elements, so it doesn't contain any sensitive information or exact personal information for which an individual's consent is to be solicited.	
APP 4: Dealing with unsolicited personal information	**Compliant**	As the *Synthetic Dataset* couldn't be rather wasn't collected using APP3, the ***DSS*** ensured that all the data was de-identified as soon as practicable (before disclosing) in *lawful & reasonable* terms as prescribed under the guidelines of *OAIC* (OAIC, 2014).	No recommendations
APP 5: Notification	**Requires Further Action or Review**	Although the *Synthetic Dataset* contains de-identified & unrecognizable (for common public) personal information yet it does contain some personal information about the individuals who may identify themselves from the 31 variables that are available at their disposal in the *Synthetic Dataset*. So, the bottom line is people are not being notified or intimated in any form whatsoever that their data is being dealt with.	The masses should have been notified and informed initially that their personal information was collected under so and so purpose before and is now is about to be published after undertaking a tremendous transformation on several variables or details for so and so purpose in the *Synthetic Dataset*. This gives the opportunity to those few people to convey their ***disapproval*** who might have some *inhibitions, reticence, or reservations,* etc against any sort of collection, use, or disclosure of their credentials in any form or manner. The aforementioned recommendations resonate with the theme of *Opt-In-*

			Processes of the **Self-Determination** *Ethical Design Principle* wherein *seeking explicit, informed consent* from those who might be affected is deemed imperative (Beard & Longstaff, 2018).
APP 6: Use or Disclosure	**Requires Further Action or Review**	The purpose of disclosing the personal information in a modified version is absolutely different from the primary purpose of collection which was invariably for assisting the people who are vulnerable and disadvantaged. Under *exemptions,* it would have been acceptable if the disclosure of information was necessary for conducting an *analysis of statistics* on **public health & safety**, but it is not the case here as the scrambled details cannot reflect the true image required for such *reporting of statistical analysis*. No *consent* is obtained anywhere nor is any *intimation* provided as discussed earlier to approximately 5 million people (some repeated) whose modified records are being disclosed for *research, teaching/learning purposes*, etc.	Based on core aspects of *Intended Purpose* and *Nudges* of the **Responsibility** *Ethical Design Principle* (Beard & Longstaff, 2018), Communicate very clearly to the people whose data or personal information is being used and represented in any form that the *Synthetic Dataset* is to be published for research, learning, and reference purposes and their personal information will not be compromised at all, then see do they consent or dissent or perhaps have any issues that need to be addressed?
APP 7: Direct Marketing	**Not Applicable**	***PIA*** is for *intended welfare policy schemes* and the *Synthetic Dataset* is deliberately created for *research, teaching & learning, reference purpose* only, and hence the distorted dataset is not available or suitable for direct marketing.	No recommendations

APP 8: Cross Border Disclosure	**Compliant**	The *Synthetic Dataset* is only published and disclosed with restricted access (to the overall ***PIA*** data) to limited people upon request. The prospective recipient can be from Australia or overseas also, but everyone has to accept the *terms and conditions* for accessing the data and this process is similar and mandatory for all. Reasonable steps are taken to ensure that the overseas recipient does not breach *APPs* or *misuse* the data as per the general *Australian Data Archive (ADA)* guidelines that one has to accept when requesting access to the synthetic data (Reddel, 2018).	No recommendations
APP 9: Government Related Identifiers	**Compliant**	The original *Personal ID* of individuals along with their partners are not disclosed as a distinctly *universal government related identifier*. *Unique Identifiers* are created for use in the *Synthetic Dataset*, but they don't uniquely identify a living entity nor mimic the original identifiers. So, this principle is being meticulously followed to conceal the true identity of the people.	No recommendations
APP 10: Quality of Personal Information	**Not Applicable**	As the information for the *Synthetic Dataset* is collected from ***PIA*** data which was originally from a reliable source containing relevant, up-to-date, accurate, and complete information, no steps were required to be taken by ***DSS*** & ***Data61*** to ensure the quality of information. Moreover, the information is deliberately altered to serve some other purpose while disclosing and	No recommendations

		this quality principle isn't applicable here as an individual isn't being directly affected by the quality of information because it is de-identified on purpose for privacy & confidentiality reasons so there is no point in focusing much on this.	
APP 11: Security	**Compliant**	As we know, the *Synthetic Dataset* contains limited de-identified information which are derived from *privacy-preserving algorithms* to secure the original identities. This is done in order to protect the privacy and confidentiality of personal information by ensuring no unauthorized access to it. Furthermore, as the synthetic data by virtue of adopted *privacy by design approach* is a highly modified & transformed data not containing any personal information that can be used to identify a living entity, it cannot be misused to leak comprehensively aesthetic personal information.	No recommendations
APP 12: Access	**Not Applicable** (when unrecognized) & **Not Able to Be Determined** (when recognized)	Since the people whose information is de-identified in the *Synthetic Dataset* are unaware of their data being manipulated (in a seemingly innocuous manner) before presenting, they aren't expected nor observed to seek access & obtain details about their existing data. However, what happens if they somehow do become aware and disturbed after discerning that their data was taken *covertly* and was later published when studying the synthetic data after requesting & then attaining access to it is something which cannot be	Prepare for all sorts of situations and render a detailed description of the procedures and protocols to be followed in different situations especially when someone identifies himself/herself in the *Synthetic Dataset* and seeks access to the relevant details for cognizance on the information being held about them. This recommendation is inspired by the **Accessibility** *Ethical*

		determined at the moment and remains to be seen (Gleeson, 2016)	*Design Principle* (Beard & Longstaff, 2018).
APP 13: Correction	**Not Applicable**	As all the personal information is highly distorted in the *Synthetic Dataset* on purpose for the safety and security of an individual's privacy, this dataset is utterly unfit for anyone to identify any existing human being. Therefore, there is no point or need of correcting any unrecognizable personal information pertaining to an individual as they are very likely to be impervious to the consequences of a publication that doesn't reveal their true identity or sensitive information.	No recommendations

REFERENCES

Beard, M., & Longstaff, S. (2018). Ethical By Design: Principles For Good Technology. *The Ethics Centre*.

Gleeson, M. (2016). 'Priority Investment': Code for attacks on most vulnerable. *Green Left Weekly*(1112), 5.

OAIC. (2014). *Guide to undertaking privacy impact assessments*. Retrieved from Office of the Australian Information Commissioner: https://www.oaic.gov.au/privacy/guidance-and-advice/guide-to-undertaking-privacy-impact-assessments/

Povey, J., Baxter, J., Ambrey, C., Kalb, G., Ribar, D., & Western, M. (2018). The power of linked data: Evaluating diverse multi-program projects designed to reduce welfare dependence.

Reddel, T. (2018). Using people-centered evidence to shape policy. Australian Institute of Family Studies [Press release]. Retrieved from https://aifs.gov.au/aifs-conference/using-people-centred-evidence-shape-policy-strategy-and-implementation

Services, D. o. S. (2019). Australian Priority Investment Approach to Welfare [Press release]. Retrieved from https://www.***DSS***.gov.au/review-of-australias-welfare-system/australian-priority-investment-approach-to-welfare

Services, D. o. S., & Data61. (2017). *Synthetic Priority Investment Approach Data*. Retrieved from ADA Dataverse: https://doi.org/10.4225/87/FASD1J

Spokes, T. (2018). Public Access to Priority Investment Approach (PIA) Data. Retrieved from https://www.communitygrants.gov.au/sites/default/files/documents/03_2018/tablebuilder_overview_brisbane_and_townsville.pdf