5 March 2024

John Fabrizio
fyi-request-25669-fe7100a4@requests.fyi.org.nz

Tēnā koe John

Official Information Act (OIA) request: IDI data retention and privacy concerns

On 8 February 2024, you contacted Stats NZ requesting, under the Official Information Act 1982 (the Act), the following information:

> *Based on my understanding of the Integrated Data Infrastructure (IDI), it appears that you possess highly sensitive information regarding every individual who has interacted with a government agency, at-least, within the past 20 years, and your intentions is to indefinitely retain all future interactions with government agencies, that is recorded by administrative data later sent to you. Statistics New Zealand frequently asserts that "it is extremely difficult to extract information about a specific individual," yet I believe this may not accurately reflect the truth.*
>
> *Each distinct identity within the IDI has its own "snz_uid." Nearly every snz_uid is linked to at least one encrypted identifier, such as "snz_ird_uid" or "snz_dia_birth_reg_uid," alongside identifiers like a birth registration number, National Health Index (NHI) number, Inland Revenue Department (IRD) number, or a physical address. While encryption may hinder direct searches for personal identifiers like NHI/IRD numbers known to researchers, the new unique number generated through encryption is still inherently tied to a single identifiable person. Furthermore, through analysis and using public databases, it's possible that most encrypted addresses could also be easily de-identified.*
>
> *Although we cannot be absolutely certain that every unique identity within the IDI, and every variable linked to that identity, definitively belongs to a specific individual, we can reasonably infer that most of it does, especially considering the permanance of NHI/NSN/IRD numbers throughout an individual's life. Notably, variables within the IDI, such as birth parents' date of birth (DOB) and individual DOB, alongside place of birth, enable the tracing of a particular identity to a snz_uid within the IDI, thus facilitating access to associated data such as mental health records, criminal records, and tax records.*
>
> *While other agencies routinely discard individual-related information to safeguard privacy, it appears that Statistics NZ indefinitely retains administrative data provided*

by cooperating agencies in the form of "de-identified" and pseudo-anonymous datasets, data that cooperating agencies themselves may securely dispose of at a later date following their own disposal schedules.

1. Do you retain the encryption key for every encrypted variable within the IDI indefinitely?

2. Is there a practice of routinely securely disposing of the encryption key for certain encrypted variables? which which put limits on the continuity of some linkages

3. Do you have plans to further enhance the confidentiality or anonymization of old IDI instances in the future? for example archived IDI data.

4. Does the value of insights gained through research using the IDI outweigh the risk of a malicious actor gaining access to a complete copy of the IDI database, thus acquiring knowledge of any individuals' interactions with most government services going back past the 1990s for many agencies? I believe that consolidating records from various government agencies into a singular database, thereby establishing a single point of failure and relying solely on "de-identification" as a means of protection, poses a significant security and privacy risk that impacts all citizens of New Zealand.

Before I address each of the four questions in your request, I would like to clarify your comments on the retention of public records, and the way encryption works in the Integrated Data Infrastructure (IDI).

Under the Data and Statistics Act 2022 and our disposal schedules, data is held for as long as it still holds administrative and statistical value. Given the IDI is a longitudinal research database, we hold the data for a long period of time. Many of the agencies that supply data to the IDI also have long-term data retention policies for their own business use.

When data is processed to form the IDI, agency identifiers are encrypted (for example, an IRD number is encrypted to the snz_ird_uid number).

Data in the IDI is linked together using probabilistic linking methods. This involves matching the records in one data source to the records we believe are the most likely associated with that person in another data source. We link records using personal information such as name, sex, date of birth, and address.

When datasets are linked together, a snz_uid is assigned. This number is only created during our linkage processes, and is not formed through direct encryption of another identifier. The snz_uid value is re-assigned each time the data is updated and re-linked, three times per year.

Once we have linked records across data sources, we remove all identifying information. The types of variables removed include:

- Unencrypted unique identifiers
- Names of people or businesses
- Day of date of birth
- Address strings (such as street number and name)

I will now address each of your questions in turn.

> *1. Do you retain the encryption key for every encrypted variable within the IDI indefinitely?*
> *2. Is there a practice of routinely securely disposing of the encryption key for certain encrypted variables? which which put limits on the continuity of some linkages*

Encryption keys are kept to enable the processing, production and security of the IDI. We retain the same encryption key for encrypted variables to ensure research can be completed consistently across time. Encryption keys are only used by the analysts and developers that work directly on the production of the IDI. Stats NZ staff members who do not work on the production of the IDI do not have access to encryption keys. We follow guidance from the Government Communications Security Bureau on encryption keys and update them as old methods of encryption become outdated.

> *3. Do you have plans to further enhance the confidentiality or anonymization of old IDI instances in the future? for example archived IDI data.*

We do not have plans to further enhance the confidentiality or anonymisation of old IDI instances.

Old instances of the IDI are not made available to researchers. These are archived within the secure Stats NZ network until no longer required. These are not accessed, even by Stats NZ staff, unless historic issues need to be investigated.

Given this, the fact they are within our secure network, and that they contain only de-identified data, we do not see a need to further enhance the confidentiality or anonymisation of old IDI instances. They are kept for record keeping purposes until they are disposed of.

> *4. Does the value of insights gained through research using the IDI outweigh the risk of a malicious actor gaining access to a complete copy of the IDI database, thus acquiring knowledge of any individuals' interactions with most government services going back past the 1990s for many agencies? I believe that consolidating records from various government agencies into a singular database, thereby establishing a single point of failure and relying solely on "de-identification" as a means of protection, poses a significant security and privacy risk that impacts all citizens of New Zealand.*

Yes, we believe that the value of insights gained through research using the IDI outweighs potential risks. The IDI is used to conduct research for the public good and to

support the design, delivery and review of public services. For example, the IDI is used for evidence-based policy development and review, informing school funding formulas, understanding the impact of the COVID-19 pandemic, research into social issues like gender pay gaps, and producing insights that inform decision making with the aim of producing greater wellbeing and equitable outcomes for Aotearoa New Zealand.

There are over 350 research projects in progress that use integrated data, with researchers from over 70 organisations including Stats NZ.

We expect researchers to publish research outputs derived from integrated data. We share these outputs on the Stats NZ storehouse (https://cdm20045.contentdm.oclc.org/digital/collection/p20045coll17). This database contains a searchable database of projects that have used integrated data since 2001, and outputs produced from the data, such as academic research papers and policy reports.

Stats NZ has conducted research into public attitudes towards integrated data, with the results published on our website: https://www.stats.govt.nz/corporate/public-attitudes-to-data-integration. One of the main findings that suggests the public supports the use of personal data in this way has been that data integration was seen as more acceptable when the data was *"completely depersonalised and anonymised"*, and strict controls were in place around access and use of the data. Given our application of the Five Safes Framework, particularly around 'safe data' and 'safe settings', we consider that we meet this expectation.

Protections

At the same time, we acknowledge the concerns about potential security and privacy risks associated with consolidating records into a singular database like the IDI. With this in mind we take a number of precautions to keep the data safe.

We work with the agency supplying the data before integration, to understand which variables in their data are uniquely identifying, and those that should not be made available to researchers for privacy reasons, and ensure these values are removed or encrypted before they are made available for research.

We provide partial date of birth or date of death fields for research, where the day of birth or death is removed, but the month and year are retained. This enables research to be conducted based on age cohorts without identifying individuals.

Similarly, all address strings are mapped to a location database, and then encrypted. The encrypted values are matched with geographic classifications such as meshblocks, allowing research based on location without revealing identifiable personal information.

Only a small number of highly trained Stats NZ staff have access to identifiable information supplied by agencies, for these staff to perform linking processes and quality assurance. Researchers using the IDI only have access to de-identified information.

Confidentiality and privacy are front of mind in all the work we do. The data in the IDI is kept safe through our application of the internationally recognised Five Safes Framework:

- Safe People – all researchers are vetted and must commit to using data safely.
- Safe Projects – all requests to use integrated data must be able to demonstrate that they are in the public interest.
- Safe Settings – data can only be accessed through a secure virtual environment, and there are a range of other access controls.
- Safe Data – data is de-identified and researchers only get access to the data they need for their project.
- Safe Output – all outputs must be checked for identifying information before they leave the secure environment to ensure the data has been confidentialised.

Before new data is added into the IDI, a robust Privacy Impact Assessment is conducted to assess the value and risks associated with each integration. These are published on the Stats NZ website. For example: https://stats.govt.nz/privacy-impact-assessments/privacy-impact-assessment-for-adding-new-zealand-health-survey-nzhs-data-to-the-idi/

In addition, there is an overarching Privacy Impact Assessment for the IDI, published on the Stats NZ website https://stats.govt.nz/privacy-impact-assessments/integrated-data-infrastructure-overarching-privacy-impact-assessment/

Should you wish to discuss this response with us, please feel free to contact Stats NZ at: OfficeoftheGSCE@stats.govt.nz.

If you are not satisfied with this response, you have the right to seek an investigation and review by the Ombudsman. Information about how to make a complaint is available at www.ombudsman.parliament.nz or 0800 802 602.

It is Stats NZ's policy to proactively release its responses to official information requests where possible. This letter, with your personal details removed, will be published on the Stats NZ website. Publishing responses creates greater openness and transparency of government decision-making and helps better inform public understanding of the reasons for decisions.

Nāku noa, nā

Mike Webb
Senior Manager – Executive & Government Relations | Office of the Chief Executive
Stats NZ Tatauranga Aotearoa
stats.govt.nz