SIA
SOCIAL INVESTMENT AGENCY
ORANGA TANGATA

Investing in what works for better lives

SIA's Beginners' Guide to the IDI

How to access and use the Integrated Data Infrastructure (IDI)

Version 1: December 2017

SIA- 2017-0255

New Zealand Government

Creative Commons Licence

Liability

Citation

Social Investment Agency 2017. *Social Investment Agency's Beginners' Guide to the Integrated Data Infrastructure*. Wellington, New Zealand.

ISBN 978-0-9951022-5-5 (online)

Published in December 2017 by and scheduled for review in May 2019
Social Investment Agency
Wellington, New Zealand

Intended audience

This paper is intended for analysts, policy teams, and managers who are interested in using the Integrated Data Infrastructure for analysis to support a social investment approach.

Resources, tools and guides

The SIA is developing a range of tools, products and guidance to enable agencies to develop their social investment approaches, and analyse and measure the impact and effectiveness of the services they're delivering.

Integrated data for social investment

This guide explains how to access and use the Integrated Data Infrastructure (IDI) for social investment analysis. When we analyse integrated data it allows us to study people's journeys over time in order to learn what works, for whom, and at what cost. In the spirit of transparent, replicable, reusable and extendable work, the Social Investment Agency (SIA) is sharing what we've learned about using this resource.

This paper:

- describes what the IDI is
- explains how to apply for access to the IDI
- how to use the IDI for your analysis, and
- how to use the analysis once it is complete.

The aim of this guide is to encourage others to make use of the IDI by providing a succinct introduction and overview.

The IDI is maintained and operated by Statistics New Zealand (Stats NZ); while this guide is intended to introduce the IDI as a resource for social investment analytics, Stats NZ should be your first point of contact for any further queries via access2microdata@stats.govt.nz or (04) 931 4253.

In order to access the IDI you will need to apply for access via the Data Labs with Stats NZ.

What is the IDI?

The IDI is a large research database curated by Stats NZ. It contains matched, de-identified data on people and households in New Zealand collected by Government agencies, Stats NZ surveys, and non-governmental organisations (NGOs). The IDI contains longitudinal data on more than nine million individuals, spanning accident compensation, crime, education, health, medical, social welfare, tax data, and others.

Stats NZ receives new data regularly and updates the IDI quarterly. When integrating new data into the IDI, Stats NZ links each individual's records across multiple datasets before removing all identifiable features such as names, NHI numbers or IRD numbers. This allows researchers to view individuals' records and interactions across services and agencies, minimising any risk of the individual being identified. This is one of the 'five safes', a framework Stats NZ uses to protect data in the IDI. Stats NZ regularly checks matched records as part of their quality assurance.

In cases where Stats NZ is unable find an individual's exact match across tables they use probabilistic linking. This process estimates the closest possible match using all available information.

The IDI can be accessed from secure environments called Data Labs. There are Data Labs at Stats NZ offices. Organisations may apply to Stats NZ to set up their own Data Lab if they intend to use the IDI for a prolonged period of time.

Why is the IDI valuable?

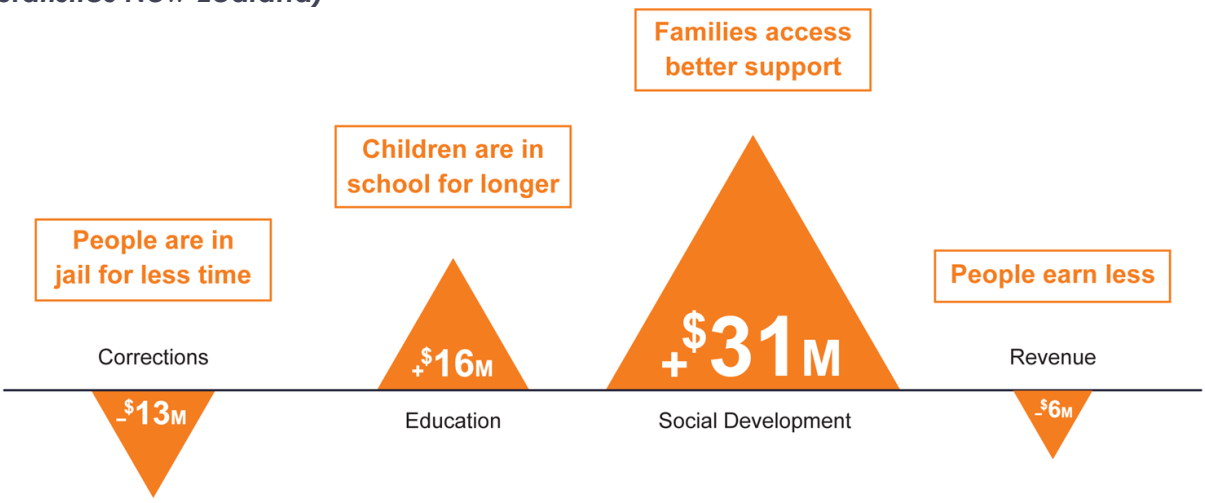
The IDI is a world-leading resource with a great breadth and depth of data. Researchers can use it to analyse populations and investigate the impact of services and programmes on people's lives.

When we use a single organisation's data to study the population they serve, we can't see much further than the outputs of the service provider and the trends indicated by annual snapshots and averages. Because the data in the IDI is matched and available at the level of the individual, it allows us to see the compounding and interrelated factors that affect peoples' lives and needs at a far more nuanced level, measure outcomes for population cohorts over time, and see outside of silos and service lines. This helps us learn what works, for whom, at what cost.

The Social Investment Agency (SIA) has produced population analyses to demonstrate the value of this sort of data analysis. One example is the Social Housing Test Case. The analysis identified all the people who applied for social housing in 2005/06, distinguished a group of successful applicants from a statistically comparable group of unsuccessful applicants, and compared their outcomes over time. The intention of this analysis was to test whether the fiscal return on government investment across the social sector could be calculated based on people's access to social housing.

The analysis quantified the fiscal return to Government of placing people in social housing. Some of the key findings are presented below in Figure 1.

Figure 1: Results of successful applicants in the Social Housing Test Case (Supplied by Statistics New Zealand)



The study also helped identify areas of potential new work, such as: investigating the balance of social and economic returns, defining measures of social well-being, and improving or understanding of which people would benefit the most from social housing.

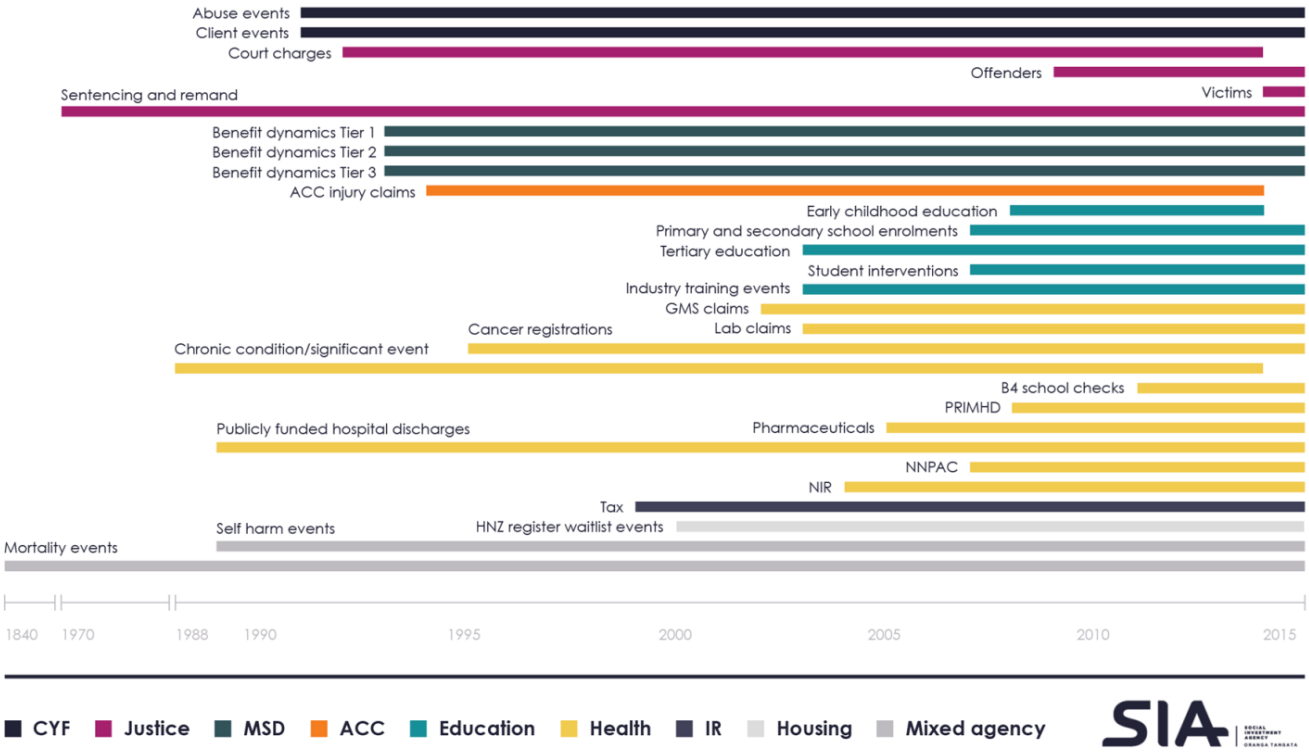
You can read more about the Social Housing Test Case here and in the technical report.

What data is in the IDI?

If you are thinking of using the IDI for an analytics project, you'll need to know what data it contains so that you can plan your analysis and apply for access. There are several ways to view and understand this using:

- the list of the types of data available in the IDI on the Stats NZ website.
- the resources developed by the SIA to assist analysts using the IDI for social investment purposes. One of these resources is the Social Investment Measurement Map (SIMM). The SIMM lists person-centred outcomes that can be measured using the Social Investment Analytical Layer (SIAL). The SIMM aims to help future analysts who are unfamiliar with the IDI data to gain a better appreciation of what is available.

Figure 2. Illustration of records available using the Social Investment Analytical Layer, and their timeframes



How do I access the IDI?

Step 1: Assess your capability and eligibility

Assemble the right team

Without the appropriate expertise it will be difficult to get the most out of the IDI.

The datasets contained within the IDI are in an inconsistent format and require somebody with the ability to apply code to arrange the datasets into an accessible state. The SIA has developed tools to reduce the amount of data manipulation required, but new users will still need somebody on their team who can code in one of SAS, SQL, STATA, or R languages. Bear in mind that because the IDI is a SQL warehouse, your team will benefit from having at least one member who is proficient in SQL.

Subject matter expertise in the relevant policy or research area is necessary to develop the scope of the research project and produce a substantive analysis from the datasets.

Check your eligibility

To use the IDI, teams must apply for access to a Data Lab. Stats NZ operates three Data Labs, one in each of its Auckland, Christchurch, and Wellington offices. Teams may apply to set up their own Data Labs for long term projects.

Stats NZ evaluates applications to access the Data Lab against very specific criteria, and will only approve access if it is satisfied that:

- the research is for a statistical purpose
- the research is for the public good
- the research will be conducted by a credible team
- suitable data is available
- Stats NZ can enforce an agreement.

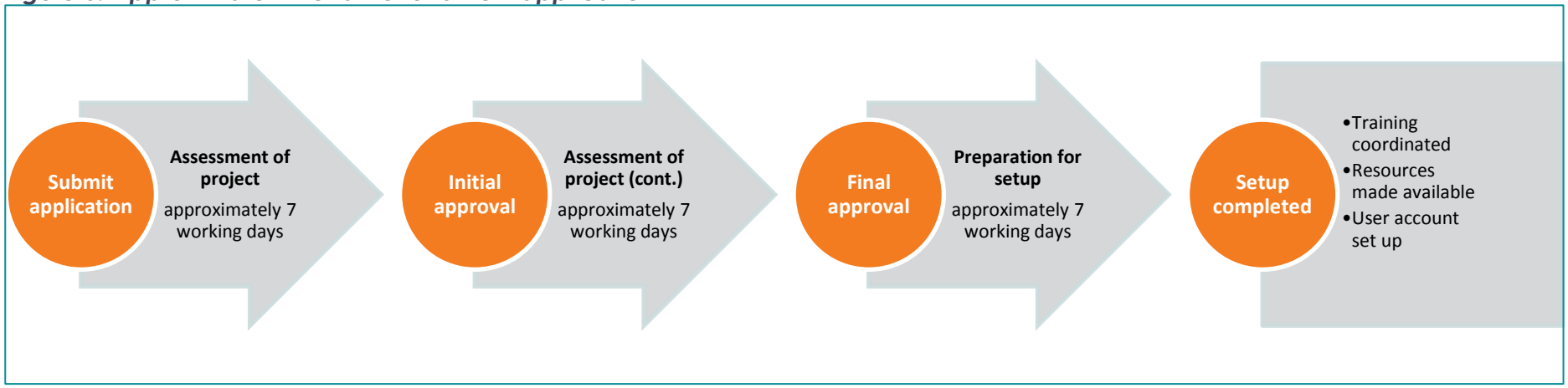
The Stats NZ website also highlights potential barriers to successful application.

Tip: To avoid delays applicants should review these barriers and consult with Stats NZ staff before submitting an application. They will be able to identify whether the data meets your projects requirements and provide advice to strengthen your application.

Stats NZ Microdata team contact details:
access2microdata@stats.govt.nz or (04) 931 4253

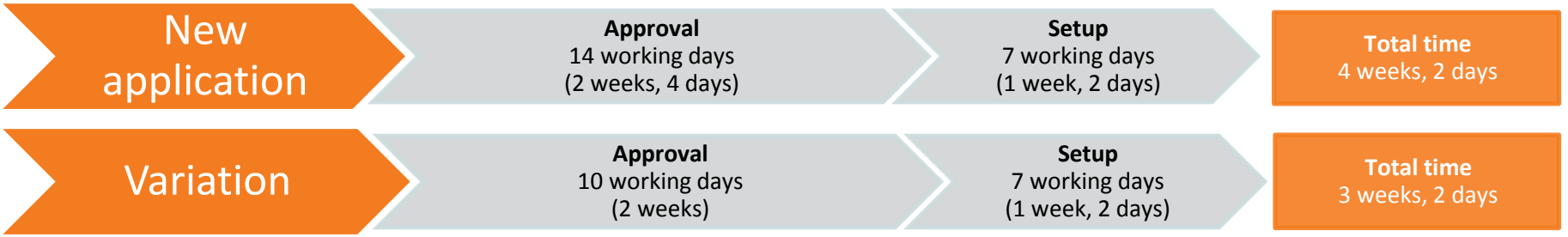
Step 2: Consider the application process and submit

Figure 3. Approximate timeframe for a new application



A new application takes approximately four weeks (21 working days). Stats NZ suggests that applicants allow one month to complete the process. Applications to alter existing Data Lab access permissions are generally approved more quickly, as below:

Figure 4. Difference between a new application and a variation



Tip: Make sure to include any data analysts, policy analysts, managers or others that may need to view outputs from the Data Lab in your applications as early as possible. Otherwise you will need to apply for their access and wait for approval later in the project.

There are three types of applications: one to access the Data Lab, one to alter an existing application, and one to create a new Data Lab.

Information required for each application type

Accessing microdata in the Stats NZ Data Lab	Varying an existing microdata project in the Stats NZ Data Lab	Setting up a remote Data Lab
▪ Your contact and organisational details ▪ Whom you want to provide access for ▪ Software requirements ▪ Research question and scope ▪ What data do you need to answer your research question? Application form	Any changes from an original application, involving: ▪ Adding/removing peoples' access to the Data Lab ▪ Variation of research question ▪ Access to different datasets ▪ Additional software ▪ Time extension Application form	▪ Your contact and organisational details ▪ Facility details ▪ Security checklist Application form

Tip: There are some restrictions on the availability of tax data.

The costs involved in an application, setup, and checking-out results

You should account for the following costs for successful approval of an application (as outlined on the Stats NZ website):

Stats NZ charges	
Assessment and approval of an application	If successful $500 No charge if unsuccessful
Output confidentiality checking	First 15 hours no charge, then $115 per hour, per project

Step 3: Getting set up in the Data Lab

Once final approval has been given, you will need to arrange an appointment with staff from the Stats NZ Microdata team to go over the necessary documentation and training. This is to ensure Stats NZ have staff available to brief analysts and provide some training on the ins and outs of the system, including the confidentiality process.

The computers in Data Labs do not have access to the internet so anything you need access should be arranged beforehand with the Stats NZ Microdata team. Your team's preferred software, coding tools or other requirements should be confirmed with Stats NZ beforehand.

What resources are available?

The size and scope of the IDI represents a challenge. First, navigating the database can be difficult because the IDI consists of more than 550 tables from 14 organizations, each with their own approach to structuring their data. Second, extracting the relevant data can be slow because the core of the IDI is 380 GB (exceeding 1 TB with additional tables), with some tables exceeding 50 million records.

You can prepare your team to use the IDI with the assistance of information repositories, dictionaries, tools and guides. Some of these resources are available outside, and others inside, the Data Lab.

It may seem daunting but there are resources to clarify and simplify the work.

Git & GitHub

Git is a version control system. GitHub is a web-based Git hosting service. GitHub can be used outside the Data Lab as a repository of complete code and reference documents.

If you are unfamiliar with Git or GitHub, you can read the SIA wiki page. Many of the repositories and wiki pages in SIA's GitHub are publically available and provide information aimed at people with coding experience about SIA tools and examples of population analysis.

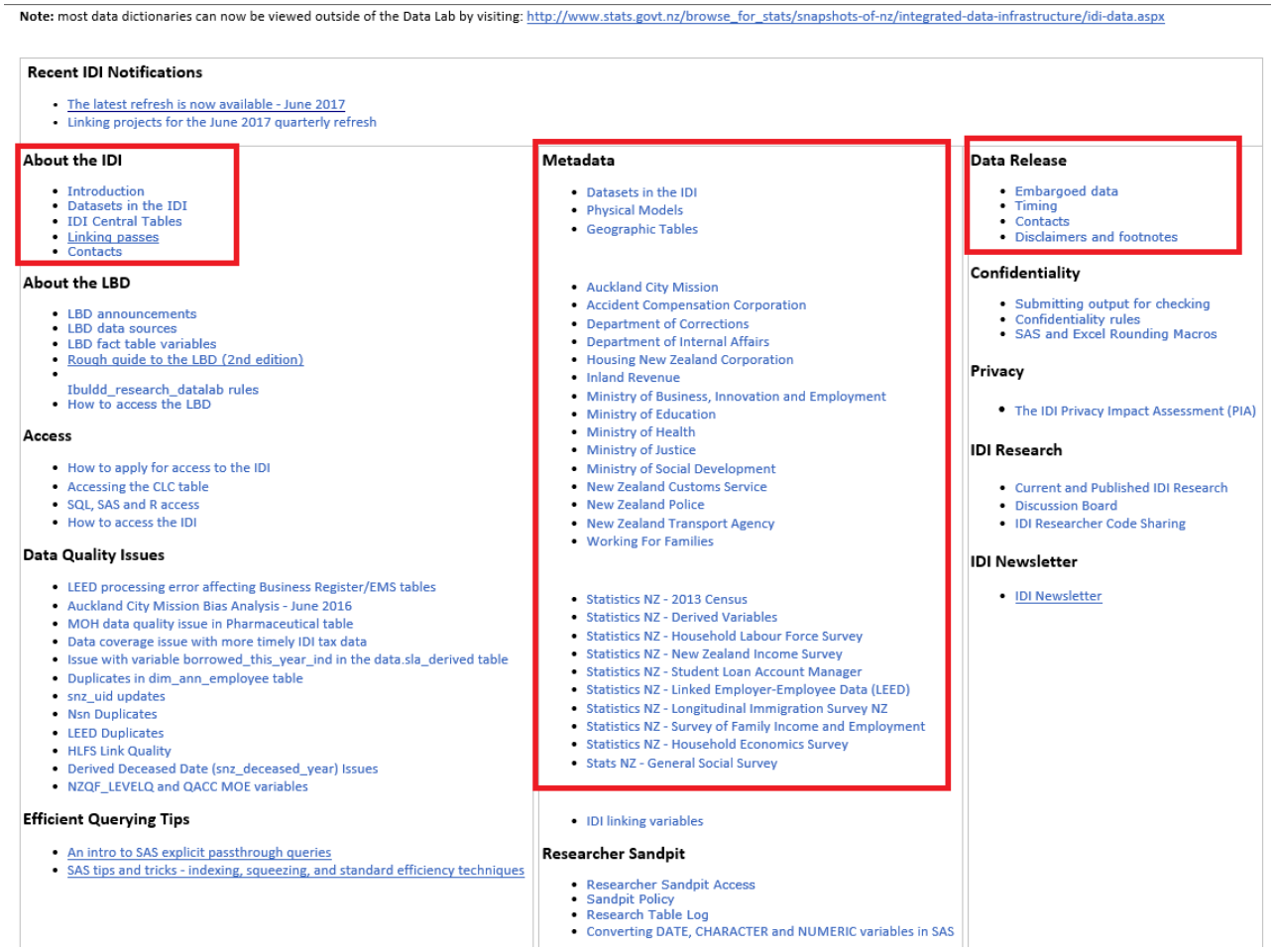
Version control is not yet available within the Data Lab but Stats NZ are in the process of implementing it.

Data lab wiki

The IDI has its own set of wiki pages, accessible only from within the Data Lab. Opening the default web browser in the Data Lab will bring you to the front page of the wiki.

Figure 5 shows the home page of the IDI wiki. Three areas that are likely to be of interest are highlighted in red: overview of the IDI, additional data dictionaries (under the heading Metadata) and confidentiality.

Figure 5: The landing page of the IDI wiki (Supplied by Statistitcs New Zealand)



Data dictionaries

As you encounter unfamiliar datasets, referring to the data dictionaries will assist you in identifying the variables of interest. Most of the tables in the IDI are in the data dictionaries and can be found on Stats NZ website.

For a small number of tables, the data dictionaries are not located on the Stats NZ website. They are instead in the Data Lab wiki under the Metadata heading.

The data dictionaries in the Data Lab contain definitions for all variables. They are listed under classification documents in the IDI wiki.

It is important because some of the datasets have a vast quantity of variables that are not immediately identifiable. For example, the Ministry of Social Development provides data on student loans and allowances which includes a column of information on the residency status of a student. The data dictionary helps to identify these variables.

Social Investment Analytic Layer (SIAL)

The datasets available in the IDI are not structured in a standard accessible way. The IDI contains more than State Sector Organisations' data tables, few of which share a consistent format.

To make it easier for analysts, the SIA has created tools which simplify the use of the IDI for social investment purposes. The SIAL is a piece of code that creates events-based tables from a selection of the data available in the IDI into a consistent format.

We recommend that all analysts working with IDI data make use of these tools to save time and effort, and avoid duplication of work. You can access the SIAL code on GitHub and can read the user guide here.

Social Investment Data Foundation (SIDF)

The tables generated by the SIAL code are one input into the SIDF code. The SIDF builds on the SIAL code to produce a dataset that is ready for analysis.

When analysing data, we are most often interested in summary measures for the specific time period of interest. The tables produced by the SIAL require filtering and need to be filtered and aggregated. The SIA built the SIDF to provide a standardised method for conducting this filtering and aggregation.

You can access the SIDF code on GitHub and can read the guide here.

How do I use the analysis once it's complete?

Releasing your work

To release your work from the IDI to anyone beyond the approved researchers listed on your microdata access application, and use it to support decision making you must submit your work to the Stats NZ for the output checking process.

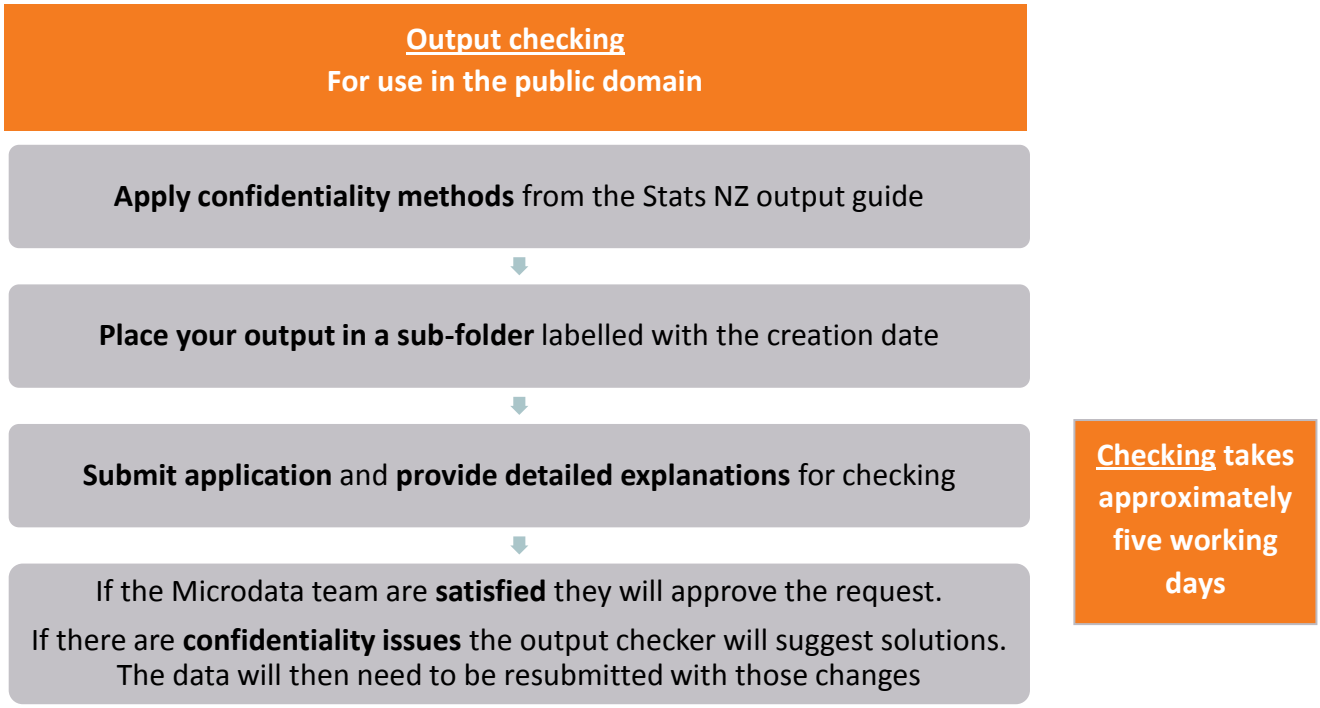
As of December 2017, there is a single process to release your work from the Data Lab for use in the public domain as outlined in figure 6.

The Stats NZ microdata output guide explains the confidentiality requirements users need to apply to their work.

This output-checking process is a mandatory requirement for using the IDI. It's how Stats NZ respects and protects the privacy of the people and households whose data is being used, by ensuring that the quality of researchers' work is satisfactory and confidentiality requirements have been applied.

The data in the IDI contains sensitive information about people, households, and businesses. For people to trust Stats NZ, and for agencies to continue supplying data for use in the IDI, it must be carefully managed and protected.

Figure 6: Overview of the process to release work from the Data Lab



Tip: Sign-out your work as early as possible to avoid bottlenecks in the output checking process.

IDI disclaimers

All research that is signed out for publication must include an acknowledgement and a disclaimer. The product must acknowledge Stats NZ as the source for any visualisation of the data and a disclaimer crediting full responsibility for the findings to the researcher.

As an example, and to meet our obligations to Stats NZ, we have included a disclaimer below. Figure 1 contains findings from the IDI and therefore requires the use of an acknowledgement and a disclaimer.

The Stats NZ Microdata output guide outlines when you need to use an acknowledgement, or a disclaimer, and what you need to say.

Disclaimer

The results in this report are not official statistics. They have been created for research purposes from the Integrated Data Infrastructure (IDI) managed by Statistics NZ. The opinions, findings, recommendations and conclusions expressed in this report are those of the author(s), not Statistics NZ or other government agencies.

Access to the anonymised data used in this study was provided by Statistics NZ in accordance with security and confidentiality provisions of the Statistics Act 1975. Only people authorised by the Statistics Act 1975 are allowed to see data about a particular person, household, business or organisation. The results in this report have been made confidential to protect these groups from identification.

Careful consideration has been given to the privacy, security and confidentiality issues associated with using administrative and survey data in the IDI. Further details can be found in the privacy impact assessment for the IDI available from www.stats.govt.nz.

The results are based in part on tax data supplied by Inland Revenue to Statistics NZ under the Tax Administration Act 1994. This tax data must only be used for statistical purposes. No individual information may be published or disclosed in any other form, nor provided to Inland Revenue for administrative or regulatory purposes.

Any person who has had access to the unit-record data has certified that they have been shown and have read and understood section 81 of the Tax Administration Act 1994, which relates to secrecy. Any discussion of data limitations or weaknesses is in the context of using the IDI for statistical purposes and is not related to the data's ability to support Inland Revenue's core operational requirements.