Working Paper
ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)

CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Ottawa, Canada, 28-30 October 2013)

Topic (iii): Modes of access to microdata

Confidentiality for integrated data

Prepared by Mike Camden, Statistics New Zealand

Confidentiality for integrated data

Mike Camden*

*Statistics New Zealand

Abstract: The Integrated Data Infrastructure (IDI) is a new feature in how Statistics New Zealand provides access to microdata. The IDI data is integrated, longitudinal, and mostly full-coverage from administrative sources. The datasets include information on people's earnings and tax, arrivals into and departures from the country, qualifications from the tertiary education system, and taking and repaying student loans. The IDI also links information from Statistics NZ's Household Labour Force Survey. Since the start of 2012, a group of researchers from across the official statistics sector has been exploring IDI data and producing results from it that would have been impossible before. In conjunction with the researchers, we developed and refined the *Confidentiality Rules for IDI*.

Integrated data raises several challenges for confidentiality. The usual concerns for utility and safety of output are heightened. Concerns for simplicity and consistency of the rules become increasingly important to researchers and output checkers. A table from the IDI could disclose particulars about individuals, their employers, or their education providers. We designed rules that protect these entities, and are as easy to apply as possible.

The dynamic environment raises more challenges: the IDI is expanding as more administrative and sample datasets are added, and the user community may become more diverse.

Traditionally, all our methods are post-tabular. In constructing our rules, we did not invent any new methods, but refined the existing ones, to maximise targeting while minimising user and checker effort.

We have provided access to microdata through our data laboratory for 15 years. We are changing this by moving towards a high-trust culture, including remote access, smoother release methods, improved documentation, and user training. We describe the challenges integrated data creates, the conceptual frameworks for handling these challenges, our confidentiality rules, and our concerns.

1 Our confidentiality rules for outputs from the Integrated Data Infrastructure

Checking output from microdata is vital for safety, but is resource intensive. Statistics NZ has built its Integrated Data Infrastructure (IDI), and the complexity of its contents and its use by researchers are growing rapidly.

As part of our strategy for ensuring safety and enabling efficient checking, we created our *Confidentiality Rules for IDI* (internal document, Statistics NZ, 2012). Researchers and checkers began using the first version of this in early 2012. We conduct internal reviews of the rules and refine them as researchers produce new outputs. In this paper we describe issues that arise from the integrated nature of the data, our frameworks for dealing with these, and our solutions.

In New Zealand, as in other countries, government and agencies within the official statistics sector want to make the best use of existing datasets. The demand for integrated data across a number of different domains is likely to continue and raises

several new issues for confidentiality of output. Our work proceeds in a changing environment.

This paper is about methods for tabular output. Integration allows us to create new forms of tables that were impossible before. The methods in our confidentiality rules are our existing ones: rounding (random and conventional) and suppression, applied in very precise ways. Our methods are all post-tabular: researchers receive anonymised but unperturbed data. Our approach to analytical output, including graphics, is explained in the *Data Lab Output Guide* (Statistics NZ, 2012).

Section 2 of this paper introduces the IDI, and the main way we currently provide microdata access for researchers: the data laboratory.

Section 3 describes the special confidentiality issues that arise because the data is integrated from many sources around a 'spine' of tax data. These issues have generated some very careful statistical thinking.

Section 4 outlines the conceptual structures we used and progressed in creating our rules.

Section 5 discusses what we have learned, and how we can apply it to the new world of open data and data hubs.

2 History of the IDI and the Statistics NZ data laboratory

Statistics NZ has been integrating data since the early 2000s. Examples of integrated datasets used for statistical outputs include the Linked Employer Employee Data (LEED), and the Student Loans and Allowances data (SLA). In 2011, we designed and built these and other collections into the prototype IDI (see figure 1). Researchers from other government agencies began using the prototype, as staff seconded to Statistics NZ, in 2012. We are considering adding more administrative and survey datasets to respond to increasing user demand.

We are well through a major shift to make microdata access via the data laboratory more researcher-friendly. We trust users' professionalism, and support them with training and clear documentation about the process and the treatment of output. The confidentiality rules form part of that. We have 'accredited researcher' status for some experienced users: they have a faster application process, and can self-release some output.

We are trialling remote access for some collections: the user's own computer becomes a thin client that enables them to interact fully with their software and the data, as if they were in our buildings. Since access to the IDI is now via the data laboratory, researchers may be able to use some of these features. See figure 2.

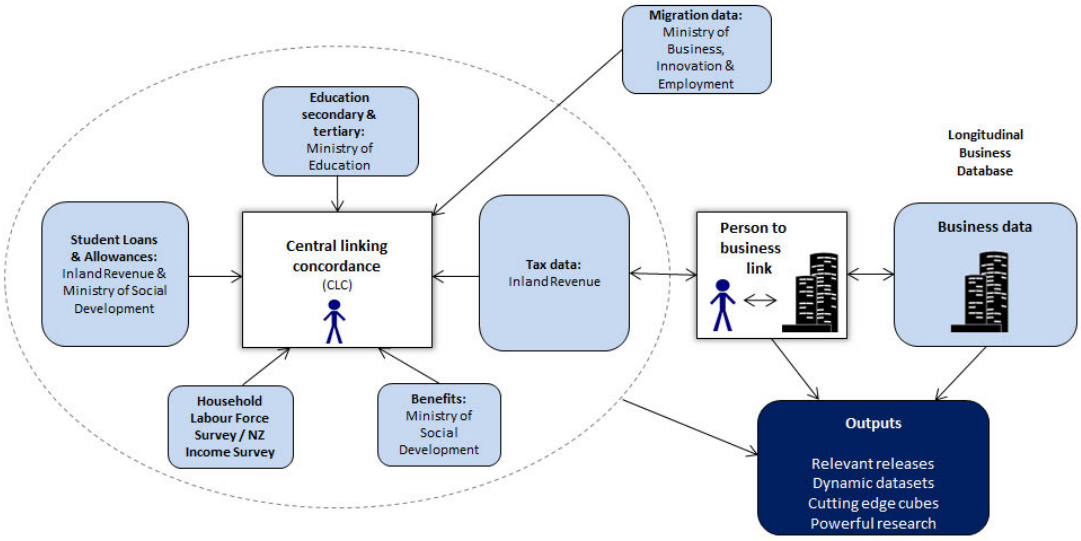


Fig 1 The IDI of 2012 had collections linked to each other, and eventually on to the tax 'spine'. The administrative collections come from outside Statistics NZ, but the Household Labour Force Survey and the New Zealand Income Survey are run by Statistics NZ and the Longitudinal Business Database contains both administrative tax data and survey data. Some links are deterministic, and others are probabilistic.

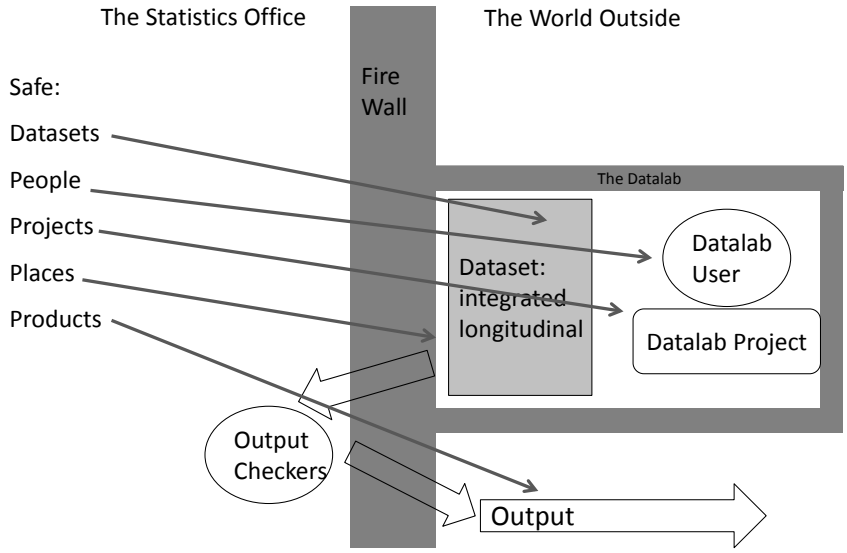


Fig 2 The data laboratory environment provides protection for IDI microdata and output via the 'five safes' framework.

3 Confidentiality issues arising from integration of data

When a statistical agency gives researchers access to integrated microdata, it needs to deal with new confidentiality issues. The researchers produce output that resembles much of the agency's own range of output, which heightens the need for consistency. The integrated data contains information about people and businesses, and the traditional distinction between confidentiality methods for these two areas needs to be rethought. Integrating education data introduces new educational entities, and these may need protection. We detail these and further issues below.

3.1 Consistency: A new aim

We began building our confidentiality rules by collecting the rules that had been made for the component collections. We found that our three aims of utility, safety, and simplicity were not enough, so added a fourth aim: consistency. Its first dimension is across the teams that can produce the output, and the second is down the component collections (see figure 3). We made evidence-based decisions about the best methods for IDI, and applied these back to the components where they were used on their own.

Suppression rules for means were consistent within collections but not between collections

	Dissemination path				Rule
Source dataset	Production	Customer requests	Datalab projects	IDI	Suppress if count less than:
Tax	Yes	Maybe	Yes	Yes	25
Student Loans	Yes	Maybe	Yes	Yes	50
Education Outcomes	Yes	Maybe	Yes	Yes	25
Benefits	Yes	Maybe	Yes	Yes	25
Migration	Yes	Maybe	Yes	Yes	2
Labour Force Survey	Yes	Maybe	Yes	Yes	5
Business Frame	Yes	Maybe	Yes	Yes	5
				IDI rule	20

Fig 3 Our rules for when to suppress averages varied across the component collections. We used evidence from applying the p% rule to settle on the rule: 'suppress where count is below 20'. Consistency sometimes needs to give way to making sure data has value and is safe.

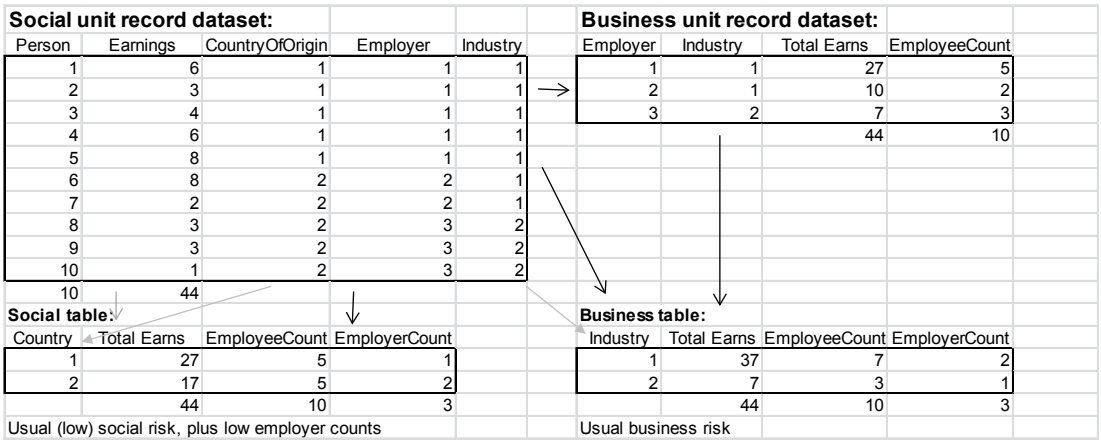
3.2 Our rules for social and business outputs from the same tax-based dataset

The IDI links people, via their tax records, to their employers. A researcher can use this data to produce social tables or business tables. They can also produce social tables with cells that contain only one or two employers (see figure 4).

Statistics NZ has different rules for social and business tables, because business tables have higher disclosure risk for the respondent. A business will occupy a cell in the dataset with competitors, and will know plenty about them already. We accept more information loss and more processing effort for business output.

We needed a definition to help users decide which set of rules to use. We consulted with researchers, and together created this definition (which is also illustrated in figure 5):

- If all the category variables that structure the table are social, then the table is social.
- If one or more of the category variables that structure the table is business, then the table is business.



Social unit record dataset:

Person	Earnings	CountryOfOrigin	Employer	Industry
1	6	1	1	1
2	3	1	1	1
3	4	1	1	1
4	6	1	1	1
5	8	1	1	1
6	8	2	2	1
7	2	2	2	1
8	3	2	3	2
9	3	2	3	2
10	1	2	3	2
10	44			

Business unit record dataset:

Employer	Industry	Total Earns	EmployeeCount
1	1	27	5
2	1	10	2
3	2	7	3
		44	10

Social table:

Country	Total Earns	EmployeeCount	EmployerCount
1	27	5	1
2	17	5	2
	44	10	3

Usual (low) social risk, plus low employer counts

Business table:

Industry	Total Earns	EmployeeCount	EmployerCount
1	37	7	2
2	7	3	1
	44	10	3

Usual business risk

Fig 4 The social dataset (top left) can produce a business dataset (top right), and also a business table (bottom right) directly. The social table (bottom left) may have a business risk hidden inside it.

Table 1
Deciding whether a table is a social or business table

	Contains business variables?	
Contains social variables?	Yes	No
Yes	Business	Social
No	Business	-

3.3 Our rules protect respondents when cells contain a small number of entities

IDI includes data about student loans, and outcomes of education. A cell can consist of people from a small number of education providers, and users of output may figure out who these providers are. Our rules require suppression of cells containing only one provider.

As we add more data collections, other entities (for example, health providers or prisons) may need protection.

3.4 Secondary suppression: When is it needed?

Our rules for tables require cells to be suppressed in some circumstances. Business magnitudes are tested with the p% rule, suppressed as needed, and protected from recalculation by secondary suppression. Magnitudes in social tables are treated more simply. For example, mean student loans are suppressed if the count is less than 20, and protected by rounding remaining values where possible. Where counts are suppressed to protect the new entities above, they are protected by randomly rounding the remaining counts. We aim to reduce researcher effort where the risk of the output allows us to do so.

Consistency in suppression processes gives way to safety and simplicity. The higher-risk business tables get stronger and more complex treatment, and the lower-risk social tables get treatment that is simpler for the researchers to apply.

3.5 Geographic and other multilevel classification variables

Several of the IDI's component datasets contain people's residential addresses. We geocode these to meshblocks, which contain an average of about 100 residents and are the smallest unit of our geographical classification. Under our new high-trust model, this meshblock data is included in the IDI where the quality of that variable allows. For other hierarchical classifications, such as the industry classification, we provide the finest level of detail that is available.

In the future, residences and workplaces may have geographical coordinates. We will need to assess the benefit and risk of providing these.

3.6 Our rules for weights when full-coverage and survey collections are integrated

The IDI at present consists mostly of administrative data from outside Statistics NZ. However, our biggest survey, the Household Labour Force Survey, is already part of it. This is a complex survey with about 28,000 respondents from our adult population of about 3 million.

The respondents all have weights that vary around a mean of 140. Researchers can produce counts, and other outputs that use the data unweighted or weighted. In both cases, we aim to protect counts that come from a small number of respondents.

The two sorts of counts require very different rules. We randomly round the unweighted counts to base 3. We conventionally round the weighted counts to a base that depends on the mean weights, and suppress below a threshold that also depends on the mean weights.

3.7 Our rules about linkage noise and disclosure risk

The IDI uses several linking processes, some deterministic and some probabilistic. Any integration stage produces some missing values, and some false matches. This unavoidable noise provides some confidentiality protection. We cannot be sure where this noise will arise in tables, so we ignore it in providing protection.

4 Conceptual structures that underlie our decisions

The ability to develop new conceptual frameworks enables agencies to respond to new information needs in a way that contributes to consistency and coherence in official statistics (Forbes and Brown, 2012). Conceptual frameworks were vital for us in creating workable output rules. Some of these appear in the figures, and two particular frameworks are detailed below.

4.1 Managing risk via the 'five safes' framework

The 'five safes', as listed by Ritchie (2006), underlie the way our data laboratory operates.

1. Safe data: we 'anonymise' by removing all source unique identifiers, names, addresses, and days of dates of death, but otherwise leave the microdata with all the richness available. IDI users receive only the datasets they need, and the concordances among them.
2. Safe projects: the application process requires a statistical and public-good purpose, and details about the data needs and outputs.
3. Safe people: the application process requires evidence of the experience and trustworthiness of the researchers.
4. Safe 'places' (physical and electronic): until recently, these have been purpose-built rooms inside our buildings, with computers that connect to the data only, and not to the outside world. We are trialling remote access, and self-release for some users.
5. Safe output: our rules provide for this – researchers apply the rules, and the checkers check they have been applied correctly.

All proposals for cross-government microdata hubs need to be analysed using this 'five safes' framework.

Risk may come from two types of people (researchers and the public), and two types of behaviour (recognition and search).

When researchers use microdata:

- they may recognise people or businesses
- they may deliberately search for one or many of them.

The first four of the five 'safes' are used to minimise these disclosure risks.

When the public use outputs from researchers:

- they may recognise people or businesses
- they may deliberately search for one or many of them.

The fifth 'safe' and our rules are used to minimise risk of public disclosure.

4.2 Our conceptual framework for the full IDI context

Researchers and checkers need to be able to quickly find their way to the rule that applies to the output. To facilitate help with this, we have developed a framework that classifies tables by:

- their content (counts versus magnitude, weighted versus unweighted)
- the entities that are the subject of the table
- table type (business versus social).

The framework can be represented as a three-way table, as shown in figure 5. This framework in also helps to clarify that we have covered all possibilities, and that we are not contradicting ourselves.

Entities to protect are:		Businesses		Persons, households		Education providers	
Table type is:		Business	Social	Business	Social	Business	Social
Output content is:	Counts:						
	unweighted						
	weighted						
	Value magnitudes						
	Count magnitudes:						
	Totals						
	Means						
	Other stats						

Fig 5 This three-dimensional conceptual framework shows: the entities we aim to protect, the table type (business or social), and the type of output we are treating. The rules fit into the body of the table.

5 Microdata confidentiality: A dynamic integration environment

Heldal (2011) describes Norway's integrated longitudinal database, the uniqueness of event histories, and the 20% sample. Our IDI is similar, but our use of the 'five safes' differs: we provide researchers with the complete anonymised datasets they require for each research project and rely on the other four 'safes' to minimise disclosure risks.

The statistical literature contains discussions about administrative data, big data, and privacy. The *Journal of Privacy and Confidentiality* devotes most of its issue 3 number 2 to administrative data. Gates (2011) recommends researching risk and public attitudes. Madans (2011) recommends education about the two aims of privacy and utility.

Chance magazine devotes a column to privacy issues. In 'O Privacy, Where Art Thou? Protecting Privacy and Confidentiality in an Era of Big Data Access', Lane (2012) states that protections are threatened as the traditional role of data producers becomes less relevant. She concludes that the brave new world of data has created new demands for infrastructures that both disseminate data and protect confidentiality.

These issues become more important when the data is integrated, anonymised, provided to researchers, and published. The literature's discussions assume these publications have been confidentialised. Our *Confidentiality Rules for IDI* are about the confidentialisation stage.

5.1 We need to make sure people understand confidentiality breaches

We have a role to make sure the public and our colleagues in other agencies understand what confidentiality breaches are, and the effort needed to prevent them. The public, as our respondents to administrative and survey data collections, understand and are rightly averse to privacy breaches. However, they may not know much about confidentiality breaches and the effort required to prevent them.

In general, a privacy breach usually consists of unit record data going outside an agency to an unintended recipient. Confidentiality breaches consist of summary output that reveals particulars about unit records.

5.2 A leadership role for the statistical institute

We developed our confidentiality rules within our Statistical Methods team, with frequent and valuable interactions with the IDI subject matter staff and the researchers. All these people are aware of the disclosure risk, and appreciate the detailed targeting in the rules. The researchers are particularly aware of the sensitivities of the educational and business entities that they work with.

Policymakers from outside this group may see building integrated data structures as having high gains in utility for little cost. The data exists already, and the structures

can be used to answer new questions. The community of experts in statistical data confidentiality has a responsibility to inform decision-makers that the plan actually has a substantial cost. This is the cost of providing safety in a detail-rich data environment, and avoiding breaches that would be even more costly. This community also has a role in showing the decision-makers how value and confidentiality can be efficiently provided.

The 'five safes' framework is an essential tool in planning an integrated data structure. It enables designers to shift the resources and streamline the processes, so they get maximum benefit without avoiding responsibility for safety.

5.3 New issues we expect to arise as access to integrated microdata expands

If we integrate more datasets and make access easier for a wider range of researchers, we will need to modify our 'five safes' strategy. The issues that we expect to arise include:

- a need for shorter and simpler projects, with users who are distanced from the data providers and less aware of their sensitivities
- a clearer definition of the distinction between statistical purpose and regulatory purpose for projects
- tables with marginal totals that have been released elsewhere
- new entities that may need protection, such as prisons, and midwifery and other health providers
- new analysis methods, from the data mining and data analytics worlds, that produce new forms of output.

We can all look forward to solving new problems as we continue to provide value and safety in this dynamic data environment.

References

Forbes, S. & Brown, D. (2012). Conceptual thinking in national statistics offices. *Statistical Journal of the IAOS* 28 89–98.

Gates, G.W. (2011). How uncertainty about privacy and confidentiality is hampering efforts to more effectively use administrative records in producing US national statistics. *Journal of Privacy and Confidentiality* 3 2 3–40.

Madans, J. (2011). Use of administrative records and the privacy-confidentiality trade-off. *Journal of Privacy and Confidentiality* 3 2 53–55.

Heldal, J. (2011). Anonymised integrated event history datasets for researchers. Proceedings: *Joint UNECE/Eurostat work session on statistical data confidentiality*.

Lane, J. (2012). O privacy, where art thou?: Protecting privacy and confidentiality in an era of big data access. *Chance* 25 4 39–41.

Ritchie, F. (2006). Access to business microdata in the UK: Dealing with the irreducible risks. *Monographs of Official Statistics – Work Session on Statistical Data Confidentiality, 2005*, UNECE/Eurostat 239–244.

Statistics NZ. (2011). *Data Lab Output Guide Version 3.0.* Statistics NZ. Retrieved 23/7/2013 from www.stats.govt.nz/tools_and_services/microdata-access/data-lab.aspx.