



**Integrated Dataset on
Student Loans and Allowances**

Privacy Impact Report

**Statistics New Zealand
Work, Knowledge and Skills
June 2007**

Liability statement

Statistics New Zealand gives no warranty that the information or data supplied in this report is error free. All care and diligence has been used, however, in processing, analysing and extracting information. Statistics New Zealand will not be liable for any loss or damage suffered by customers consequent upon the use, directly or indirectly, of information in this report.

Reproduction of material

Any table or other material published in this report may be reproduced and published without further licence, provided that it does not purport to be published under government authority and that acknowledgement is made of this source.

Published in December 2007 by

Statistics New Zealand
Tatauranga Aotearoa
Wellington, New Zealand

ISBN 978-0-478-26997-0 (online)
ISBN 978-0-478-26998-7 (print)

Contents

1	Introduction.....	1
1.1	Overview.....	1
2	Legal context / background.....	1
3	Project description	3
3.1	Official statistics and research	5
3.2	Information flows.....	6
3.2.1	Raw data server.....	8
3.2.2	Clean data server.....	9
3.2.3	Application development and testing.....	10
4	Privacy analysis	10
4.1	Source data.....	10
4.2	SLAM data.....	15
4.3	Data integration.....	17
4.4	Use and dissemination	19
4.4.1	Aggregated output tables	19
4.4.2	Data laboratory.....	20
4.5	Storage and retention	21
4.5.1	Protection against loss.....	21
4.5.2	Unauthorised access, use, modification or disclosure	22
4.6	Summary.....	23
5	Additional privacy-enhancing responses	24
6	Future compliance mechanisms	24
7	Conclusion	24
	Appendix 1: Data integration policy.....	26
	Appendix 2: Microdata access protocols	26
	Appendix 3: Privacy checklist	26
	Glossary	28
	References.....	29

1 Introduction

1. The following privacy impact report has been prepared for the Integrated Dataset on Student Loans and Allowances in accordance with Statistics New Zealand's data integration policy.
2. The information created by the integrated dataset is in the form of anonymised unit-record data that is used to produce official statistics and support statistical research, which, in particular, is needed to calculate the fair value of the student loan debt. No information that would allow the identification of any individual is published from the dataset. This report identifies the privacy risks associated with the dataset and outlines the processes for managing these risks.

1.1 Overview

3. The report begins with a description of the legal context within which the work is undertaken. It then outlines the major information flows that are required to undertake the integration process. Issues associated with the source data's integration, storage and use are then discussed, including a description of the release and access practices employed. This discussion includes an assessment of the risks inherent in these processes and the privacy enhancement and risk management procedures undertaken to mitigate them. The report concludes with a brief discussion of the compliance mechanisms to be used to manage the integrated dataset in the future.

2 Legal context / background

4. The integration of data can provide improved or new official statistics for government, businesses, communities and citizens. Integration can also avoid the need to collect additional information. However, an integrated dataset will increase the information available about an individual or business beyond that which they might be aware of. There may also be other privacy risks associated with data integration projects, such as the security of stored data. Careful management of the information and of the linking processes is required, to ensure that privacy is protected and the security of the information is maintained.
5. In the mid-to-late 1990s, submissions to Cabinet pointed to privacy issues that were present in data integration initiatives, with the need for public confidence and acceptance as concerns to be managed. In order to address the privacy concerns and risks inherent in statistical integration work, Cabinet, in its decision [CAB (97) M 31/14], assigned responsibility to Statistics New Zealand for cross-agency statistical integration work,

using information collected for unrelated purposes. Statistics New Zealand was also assigned custodianship of the resulting datasets.

6. Together, the Statistics Act 1975 and the Privacy Act 1993 form a framework to protect information about individuals and other entities when used for statistical or research purposes. Compliance with this framework ensures that the privacy concerns of individuals and businesses are not set aside for the *potential value/perceived benefits* of record linking. The Privacy Act 1993 provides protection for living natural persons, while the Statistics Act 1975 has wider security coverage, including natural or deceased persons, businesses or other entities that might have confidences to protect. Other legislation such as the Tax Administration Act, under which the Commissioner of Inland Revenue is authorised to communicate information to Statistics New Zealand for statistical purposes permitted under the Statistics Act 1975, also provides for the protection of privacy.

7. The Statistics Act 1975 and the Privacy Act 1993 are based on openness and transparency in the use of information to maintain public trust and the continued cooperation of information providers.

8. The Statistics Act 1975 regulates the collection of information, whether from statistical surveys or administrative records, for use in producing official statistics by government departments. The Statistics Act 1975 includes strict provisions to protect the security of collected information and to prevent the release of identifiable information about an individual or business. (Identifiable information includes information where identifiers have been removed, but where a third party could identify that the information relates to a particular person or business.) In accordance with the Statistics Act 1975, Statistics New Zealand ensures that information furnished to the department is kept securely, access is restricted and any publication avoids disclosing identifiable information. All data held by Statistics New Zealand is subject to these strict provisions.

9. The Privacy Act 1993 provides protection to information about an individual and applies to every agency that deals with personal information. The 12 information privacy principles (IPPs) in the Privacy Act 1993 provide the foundation upon which the Privacy Act 1993 controls the collection, use, disclosure, and storage of, and access to, personal information. Under the terms of the Act, the Privacy Commissioner must retain the capacity to conduct an independent review in the event of a complaint. As a consequence, the Privacy Commissioner is not able to approve proposals in advance. However, the Privacy Commissioner is able to signal any practices that are not permitted under the Act or that might pose the problem of perceived privacy risk. Statistics New Zealand has

taken the position that any such concerns, even of perception, should be addressed in an appropriate and defensible manner before integration commences, the intention being that Statistics New Zealand takes all necessary steps to comply with both the spirit and letter of the Act.

10. The student loans integrated dataset, initiated in 2002, is the longest running data integration project at Statistics New Zealand. It has assisted, in particular, with measures of the fair value of the student loan debt in order to meet government financial reporting requirements. Consideration of privacy risks and consultation with the Privacy Commissioner, have been ongoing throughout the early years of the project. The major redevelopments undertaken, including the extension of the dataset to cover student allowances, are a significant watershed in delivering a system in which all known privacy concerns have been addressed.

11. With all data integration projects, some form of identifying information is required to ensure that the linkage is effective. While a variety of identifying information can be used for linking, the use of unique identifiers generally provides for the most accurate linkage. This means that particular care is needed to comply with IPP 12, which prevents an agency from assigning a unique identifier used by another agency. Statistics New Zealand has worked in conjunction with the Privacy Commissioner to ensure that its data integration work complies with this IPP.

12. Existing Statistics New Zealand policies and protocols on data integration (appendix 1) and microdata access (appendix 2) are designed to protect the confidentiality of individual information, and will mitigate many of the privacy issues raised by the process of integration.

3 Project description

13. The Student Loan Scheme began in 1992. Under this scheme, students can borrow money for the payment of tertiary fees, for course-related costs, and to subsidise their living expenses. Several agencies are involved in managing the scheme and each has developed information systems to support its role, but these systems are not connected. No system contains all of the important information on borrowers – their tertiary education, their borrowing, loan balances and repayments – that is needed to analyse the key outcomes of the loan scheme.

14. In June 2000, the Auditor-General released a report *Student Loan Scheme – Publicly Available Accountability Information* that proposed Statistics New Zealand integrate selected datasets relating to the Student Loan Scheme, with a view to providing statistics

for strategic policy, financial risk management, financial reporting and forecasting. After a period of development work, including investigation of the privacy, logistical and data issues, the first integrated dataset on student loan borrowers was constructed in 2002.

Subsequently, the dataset has been updated each year since 2004.

15. From the latest update in April 2007, student allowance data has been included in the integrated dataset. The Student Allowances Scheme provides a subsidy for the living costs of students who meet certain income and age criteria. Evaluation of the integrated dataset in consultation with users showed that the ability to provide statistics on student allowance recipients would increase its usefulness substantially. At the same time, other major changes have been in progress, including re-engineering of the computer platform, redesign of the system architecture, redevelopment of the statistical production process (ongoing during 2007), and redevelopment of the annual statistical releases.

16. The following description focuses primarily on the process as at April 2007, that is, including both the student loans and allowances data, the move to a new computer platform, to a new system architecture, and to revised statistical outputs. The prospects for some specific developments that may occur subsequent to this are also noted.

17. The process of creating the integrated dataset involves collecting information on individuals from several government agencies including:

- data relating to student loans and allowances paid from StudyLink (a service of the Ministry of Social Development) from 1999 onwards
- student loan debt and repayment data from Inland Revenue for those who have received a loan or an allowance from 1992 onwards, as well as income data from 1997 onwards
- tertiary enrolment and completion data from the Ministry of Education for all formal students from 1997 onwards
- data on student loan entitlements and drawings from the Student Loan Accounts Manager (now defunct) from 1992 to 2000.

18. The data from different agencies are linked by Statistics New Zealand at unit-record level. This enables the educational characteristics of individuals who have received loans or allowances to be associated with the details of that financial support. Records are also linked longitudinally to create a record, of someone's borrowing, debt, repayment, allowance and income history that is linked to a record of their tertiary education.

19. The outputs of the process include official statistics at aggregate level, which are published in an annual Hot Off The Press release (an electronic version is made available on Statistics New Zealand's website), and a database of anonymised unit-record data held

in Statistics New Zealand's data laboratory, which is used for policy research and financial analysis. Access to this data is strictly controlled (see section 4.4.2 below).

20. Statistics New Zealand aims to update the integrated dataset annually. Each time this occurs, the dataset is completely recreated. StudyLink, Inland Revenue and the Ministry of Education provide current and back data, and the annual 'reload' covers all years back to the start of the student loan scheme in 1992 (education and income data, however, are only provided from 1997 onwards). The Student Loan Accounts Manager (SLAM) operated under contract to the Ministry of Education from 1992 to 1999. In 2000, student loan scheme account management was transferred to the Department of Work and Income (becoming StudyLink). SLAM continued to manage some loan accounts for a time during the transition, but is no longer involved in the delivery of the scheme. Its data is held permanently by Statistics New Zealand for the purpose of recreating the integrated dataset each year.

3.1 Official statistics and research

21. For the different student characteristics available in the dataset, for example, sex, age, ethnicity, field of study, level of study, and provider type, official statistics can show the average:

- amount and type of financial support in a year
- loan balance accumulated throughout study
- proportion of borrowers who have repaid in full
- loan balance by residency status, that is, in New Zealand or overseas
- income differentials for subgroups post-study.

22. The statistical releases published to date are available on Statistics New Zealand's website (<http://www.stats.govt.nz/products-and-services/info-releases/student-loans-allowances.htm>). Earlier releases focus on the number of borrowers by loan size and average borrowing/debt levels broken down by sex, age and ethnic group, as well as a number of education-related characteristics, including field of study, level of study and provider type. The statistics also include a comparison of average debt levels between those borrowers residing in New Zealand and those who had advised Inland Revenue they were living overseas.

23. The latest release *Student Loans and Allowances: 2005* is broader, incorporating statistics on financial support received through both the student allowance and student loan schemes that include enhanced time series information. It introduces several new focal points, particularly in the area of outcomes, including debt levels on leaving study

and five years after leaving study, as well as the income levels of student loan borrowers and allowance recipients one and five years after leaving study. These statistics will be updated in future releases.

24. The users undertaking policy research and financial analysis on the integrated dataset to date have been from the Ministry of Education and auditors working on contract to the Ministry of Education.

25. The work of the auditors relates to estimating the fair value of the student loan debt for government financial reporting purposes. The fair value is calculated using a model constructed for the Ministry of Education, which requires the integrated dataset to combine students' educational and demographic information with data on loans and income.

26. Policy-related research has explored such topics as:

- factors that may influence full repayment of student loans, for example, sex and qualification completion
- who makes no progress toward loan repayment
- student loans and Māori students
- the relationships between debt, qualifications and post-study income.

27. Several reports can be found on the Education Counts website.

(<http://educationcounts.edcentre.govt.nz/publications/tertiary/indexDate.html>)

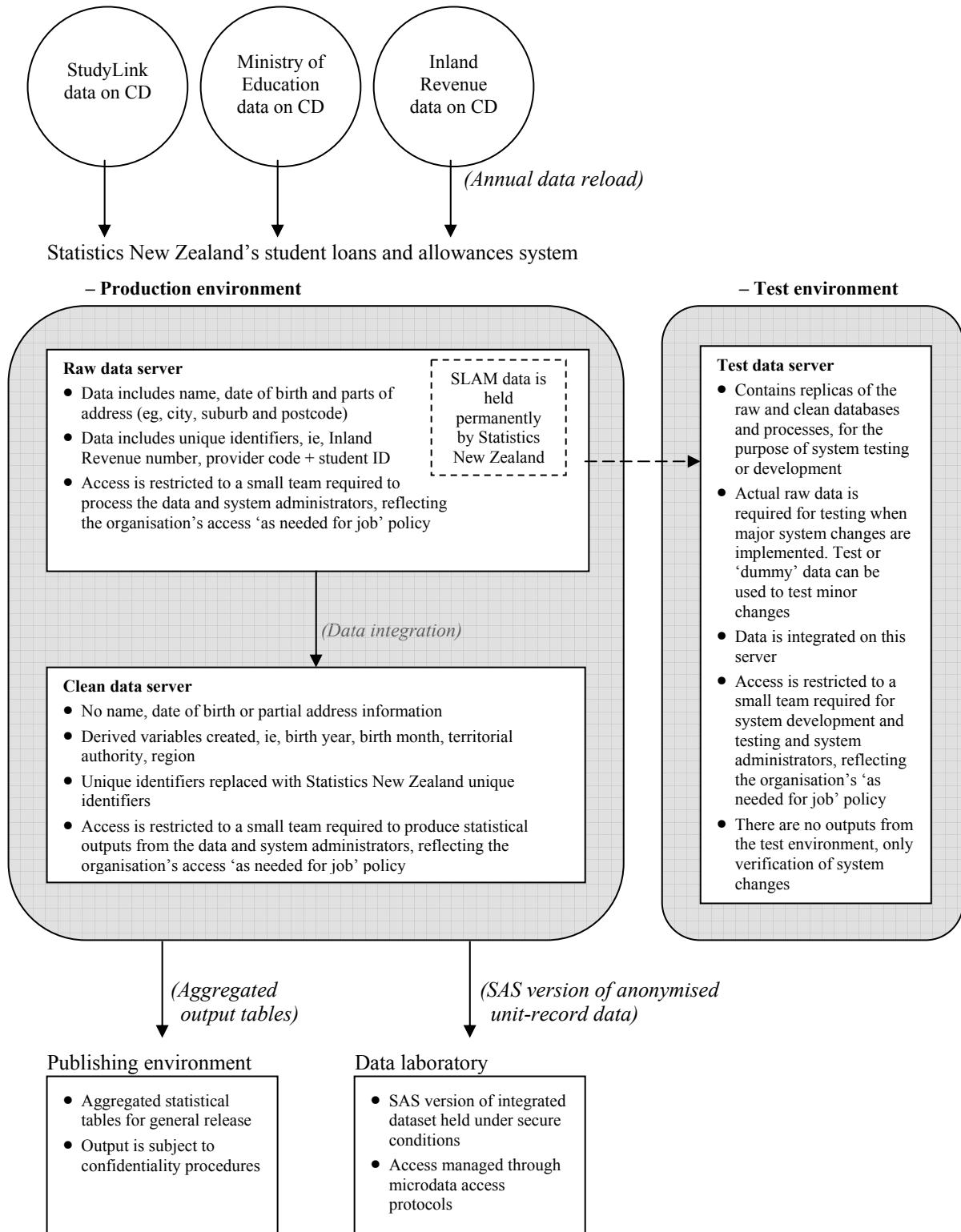
3.2 Information flows

28. Files on CD will be received by secure person transport from StudyLink, Inland Revenue and the Ministry of Education from 2007/08. (Data for the next – 2008 – release will be received late in 2007.) On arrival they are copied to a folder on the student loans 'raw' data server. The original CDs are held securely until the data has been loaded and checked, and are then destroyed. Historical data from the Student Loans Account Manager is held permanently on this server, and the original CDs have also been retained.

29. The student loans and allowances data is stored and processed in its own environment. In the *production* environment, there are two separate servers. One for the original or raw data provided by source agencies, and another for the integrated dataset. This keeps the integrated data apart from other data. A third, separate, server is used to develop and test system changes before they are implemented in production. Each server is made accessible to the smallest number of Statistics New Zealand employees practicable.

30. The following diagram shows the flow of personal information within the process.

Flow of Personal Information within the Student Loans and Allowances Data Integration Process



3.2.1 Raw data server

31. On the raw data server, a range of tests is applied to the data files received from supply agencies, to confirm that mandatory fields contain data and to check for errors in:

- the type and format of the data
- the range of values
- the number of fields and format of field names.

32. Once validated, the source data is loaded into database tables. At this stage unique identifiers, such as Inland Revenue number, and personally identifying information, such as name and date of birth, are present. In the process of being prepared for integration the data pass through a series of transformations.

- Unique identifiers are encrypted and replaced. Currently two of the data items used to match student records are considered unique identifiers. One is the Inland Revenue number included on the StudyLink, Inland Revenue and some Ministry of Education records. The other is the student ID assigned by the provider that is included on the StudyLink and Ministry of Education records (this is used in combination with the education provider code to create a composite identifier). Encryption is achieved using a Statistics New Zealand standard encryption algorithm, which derives an encrypted ‘string’ from the original unique identifier. The encrypted string is mapped to a Statistics New Zealand unique identifier which replaces the original unique identifier. The mapping table is a permanent table used each year in the student loans and allowances data integration process.
- The date of birth is used to derive two new variables – one for birth year and the other for birth month.
- Post codes are also used to derive additional variables, including territorial authority and region.

33. The original raw data, derived variables, encrypted strings and the Statistics New Zealand assigned identifiers all reside in database tables on the raw data server at this point. To date, access has been restricted to a small team of analysts, developers and testers (10 or so) who work directly on the processing of student loans and allowances data, plus five database administrators who need administrator rights to administer the systems. Once the current system developments have been implemented, and the data integration process becomes ‘business as usual’, developers and testers will only have access for limited periods if required to undertake specific work, for example, fixes and enhancements.

34. The data integration process uses specialised software that extracts the variables to be used in linking from the database tables, processes them on its own server (equally secure) and writes back the 'links' (that is, mapping information that specifies the records to be joined) to the raw data server. A database application is then used to extract records from the database tables and join them according to this mapping information, at which point they become linked anonymised unit-record data. Student records are linked within the Ministry of Education data, and then across all data sources. The linked records are written to the 'clean' data server and the integration process is complete. There is no linked data stored on the raw data server. (The links currently remain on the raw data server, and this procedure is to be reviewed during 2007.)

3.2.2 Clean data server

35. The integrated dataset resides on a separate server from data in the preceding steps. This server is designated the 'clean' data server. The data that moves through to the clean data server does not include name, address, or date of birth information, and also does not include the original unique identifiers or the encrypted strings. It contains instead the derived variables and Statistics New Zealand assigned unique identifier. (In the case of the student ID, this is first combined with the education provider code to form a composite identifier that is replaced, and neither it nor the original student ID moves into the integrated dataset. The provider code does flow into the dataset.)

36. Both the raw and the clean data servers are dedicated exclusively to the student loans and allowances data integration process. No other work is undertaken, or data stored, on these servers.

37. Statistics are released as part of each annual 'reload' of the dataset. The integrated dataset is used to create aggregated output tables, which are extracted for publication. These are written into formatted Excel spreadsheets and move off the student loans server. In the publishing environment, commentary is added and any potential confidentiality issues, for example small cell sizes, are checked and addressed as a standard procedure. Once all necessary confidentialisation measures have been applied, the results are released as a Hot Off The Press publication, and an electronic version is made available on Statistics New Zealand's website.

38. A SAS version of the anonymised unit-record data in the integrated dataset is held under secure conditions in the data laboratory, where access for research purposes is managed through the microdata access protocols.

3.2.3 Application development and testing

39. In addition to the information flows noted previously, a ‘test’ data server is also used. This is a third, separate, server that is used to test the effects of a system change on the databases. For major changes this requires copies of the actual source data. For minor changes test or ‘dummy’ data may be used. Once it is confirmed that a change will not have any harmful effects, the change can be implemented on the raw or clean data servers as required (and any copies of actual source data used are removed from the test data server).

4 Privacy analysis

40. This section describes the privacy issues associated with the use of personal information within the Integrated Dataset on Student Loans and Allowances. Compliance with Statistics New Zealand’s data integration policy will mitigate many of the risks identified in this section. However, in the interests of transparency, the underlying risks have been described, so that the extent of their mitigation by generic processes and specific actions can be assessed. The analysis looks at issues related to the source data, integration, use and dissemination, storage, and retention.

4.1 Source data

41. The data brought together in this process is supplied by several agencies involved in administering the student allowances and student loan schemes:

- StudyLink is responsible for the administration and delivery of student allowances and of student loans to students during the year of study. After the end of each academic year, StudyLink transfers those loans to Inland Revenue for collection.
- Inland Revenue is responsible for collecting student loan repayments from the year after the money is borrowed and administering interest write-offs.
- The Ministry of Education is responsible for the development of policy to support participation in tertiary education. The student allowances and student loan schemes are the main initiatives in this area.

42. The authorisation for Statistics New Zealand to obtain the information annually from these agencies comes from a memorandum of understanding (MoU) signed by the chief executive of each organisation.

43. No data is collected directly from individuals in this process. All data supplied to Statistics New Zealand results from administrative requirements within the student allowances and student loan schemes, the tax system (income data), or the education sector (enrolment and completions data). Aggregated statistics are published from each of

these databases. The supply of data for the purpose of creating an integrated dataset improves and extends the statistics available on the student allowances and student loan schemes. The data consists of identified unit records.

44. StudyLink supplies information on student loans and allowances recipients for each year from 1999 onwards. The data items include:

- name, date of birth, sex, ethnicity (voluntary question), education provider(s), student IDs, Inland Revenue number, social welfare number (SWN), city of study address, locality of study address, postcode of study address, city of postal address, locality of postal address, postcode of postal address, iwi (voluntary question), study start date, study end date, and residential status
- type and amount of loan payments, sum of repayments and refunds, loan balance and interest transferred to Inland Revenue, loan balance and interest untransferred, and whether in prison
- type and amount of allowance paid, number of weeks of allowance, student's income, number of partners, number of studying partners, total partner income, each parent's income (if parent-income tested), and NSCED (see Glossary)
- type and amount of other payments, including the accommodation benefit, A and B Bursary, Step Up scholarship, Bonded Merit scholarship and top scholar award.

45. Inland Revenue supplies information on each student loan or allowance recipient for each year from 1992 onwards, including:

- name, date of birth, education provider code, Inland Revenue number, full- or part-time student status, and residency status
- loan registration start and end dates and reason for registration end, type of identification used when registering for a student loan (for example, birth certificate, passport) and identification number
- principal transferred from StudyLink, interest transferred, interest eligible for write-off, loan balance, interest compounded, interest written off
- details of student loan repayments (including overdue repayments, capital write-off, capitalisation and penalties for the current period and overall, the current rate of repayment assessed, and the net sum of all credit/debit transfers and refunds).

46. Inland Revenue also supplies address information on each student loan or allowance recipient from the earliest student loan registration or earliest student allowance:

- aspects of main location and postal addresses including city, suburb and postcode.

47. Inland Revenue also supplies income-related information on student loan or allowance recipients for each year from 1997 onwards, including:

- gross earnings (from each employer if more than one) including salary/wages, benefits, New Zealand Superannuation, and withholding payments, as well as tax paid, employer industry code, employer's Inland Revenue number, and employment stop date

Additionally, depending on the student and allowance recipient's tax situation, information may be collected on:

- gross interest, gross dividend, estate/trust income, overseas income, partnership income, shareholder salary, rents, self-employed income, other income, total expenses claimed, taxable income, tax on taxable income, total rebates, family assistance entitlement and employee industry code.

48. For each formal tertiary student for each year from 1997 on, the Ministry of Education supplies:

- name, date of birth, sex, ethnicity, Inland Revenue number (voluntary question), country of citizenship, disability flag and TSEC_SN student number (see Glossary)
- enrolments, including education provider, student ID, qualification code, attendance – intramural/extramural, equivalent full-time student (EFTS) for this qualification for this year only, total EFTS for this student this year for all qualifications, total EFTS required for this qualification, field of study, programme start – this year, first year at tertiary institution, source of funding, ISCED level (see Glossary), full- or part-time status, programme end – this year, New Zealand Standard Classification of Education (NZSCED) field of study, qualification award category code, subfield of study and provider type
- completions, including education provider, student ID, qualification code, subject major/s, qualification award category code
- educational background, including last secondary school attended, 2005 TFEA decile (see Glossary) of last secondary school attended, regional council of last school attended, highest secondary qualification, last year at secondary school and main activity on 1 October previous year.

49. Any person who has had a student loan or allowance since 1992 will be on the dataset, and all of their data (as specified above) will continue to be supplied for up to 20 years after full repayment of the loan or last payment of student allowance, whichever is more recent. This is to allow for statistics on the employment outcomes of student loan and allowance recipients. The data supply period will be reviewed by Statistics New

Zealand, pending investigation of the long-term effects of education on employment outcomes.

50. The potential risks to privacy are greatest at the raw data stage, when the unit-record data still contains identifying variables, such as name, date of birth and address information. Access at this stage may allow the viewer to see information on an identified individual's student loan debt, income, and other personal details. It is unavoidable that some Statistics New Zealand staff members will need access to the raw data for their work on the process. However, wider access to the data than this would unnecessarily impact on people's privacy. A breach of security or confidentiality could also negatively affect public perceptions of data integration processes and the source data collections.

51. For these reasons, access is strictly controlled. The raw data is processed on a separate, dedicated server. Individuals are only granted access rights to this as necessary for the process, and are subject to role-based access control for all read and write activities. Statistics New Zealand employees who have access to unit-record information provided by Inland Revenue must sign an Inland Revenue certificate of secrecy.

52. Different items of identifying information assist in validating the data, deriving key variables, such as birth month and birth year, and in linking records from different sources. However, personally identifying information is not passed through to the integrated dataset, and does not move through to the clean data server. It is contained on the raw data server.

53. Statistics New Zealand does not contact individuals to tell them that it has obtained their personal information for use in data integration. The annual release of data from the integrated dataset published on its website provides an example of the data contained in the dataset and an outline of the process. Consultation has been undertaken with stakeholders on the recent developments to the process.

54. A description of the uses of the integrated dataset has also been provided on the Ministry of Education website since 2002, and this information has been widely circulated in the tertiary education sector. The Ministry of Education data is collected from tertiary education providers, who in turn collect it from students. The Ministry of Education provides a generic enrolment form for use by providers that indicates to students how the information will be used. Use of their information includes, among other things, to furnish statistical returns to the Ministry of Education, and students are asked to sign the form, therefore authorising the supply.

55. Inland Revenue's guide to making student loan repayments states that Inland Revenue has an agreement to supply information to Statistics New Zealand for statistical purposes. This caveat also appears in the notes to both the *Personal Tax Summary* and *IR3* form, from which some of the data comes. It is not compulsory for student loan or allowance applicants to give StudyLink the information that it asks for, but applicants are informed that if they fail to do so their application may be declined. The exceptions are for information on ethnicity and iwi, which is purely voluntary.

56. Special sensitivities may surround some of the data items collected for some individuals. These variables include:

- whether in prison
- number of partners (for example, married or de facto)
- disability
- citizenship
- ethnicity
- post-study income
- income received from benefits
- student loan registrations ended due to bankruptcy.

57. Three unique identifiers are used in the student loans and allowances data integration process. These are:

- Inland Revenue number assigned by Inland Revenue. This is used in exact matching of StudyLink to Inland Revenue data.
- Student ID assigned by an education provider. This is used in combination with the provider code in probabilistic matching of StudyLink and Inland Revenue data to Ministry of Education data.
- SWN number assigned by the Ministry of Social Development. This is used as a reference number in cases where it is necessary to back check details with original StudyLink records during the initial data validation process. It is not used to match student records but will move into the integrated dataset in future, after being encrypted and replaced by a Statistics New Zealand unique identifier.

58. The ID number of the identification used during student loan registration, for example birth certificate or passport number, is also supplied by Inland Revenue. This does not appear to have a substantive use and does not flow into the integrated dataset. Receiving this variable may assist in corroborating that reasonable steps have been taken to establish a person's identity before a unique identifier is assigned, as required by Information Privacy Principle 12(3).

59. In future, the National Student Number (NSN) may also be supplied by the Ministry of Education. Access to and the use of NSN is governed by Part 30 (sections 341 to 347) of the Education Act 1989. The use of the NSN in the integrated dataset would be a permitted use under the Act. However, access to the NSN by Statistics New Zealand would require a regulation to be made by the Governor General as an order in council. The Act specifies that any such regulation will be made only following consultation with the Privacy Commissioner. The Ministry of Education intends to seek this access for Statistics New Zealand during 2007. The approach to be taken will be drawn from the current privacy code of practice for the post-compulsory education identifier. The use of the NSN in matching would supplement the student IDs currently used in data integration. Student IDs are assigned by the education providers and lack consistency when a student has been enrolled with more than one provider, simultaneously or over time. A similar approach would be taken to encrypting and replacing the NSN as is used with the other unique identifiers outlined above. Currently the Ministry of Education supplies the TSEC_SN student number that has been created as an identifier – and has been assigned with reference to the NSN for 2003 and later years, that is, since the introduction of the NSN. This will also be used for matching in future.

60. For those in employment, the Inland Revenue number of the employer is also received. Under the Statistics Act 1975 the confidentiality provisions extend to businesses as well as individuals. The employer's Inland Revenue number is replaced with a Statistics New Zealand assigned identifier, although this procedure does not involve the encryption process as outlined for an individual's Inland Revenue number.

4.2 SLAM data

61. In addition to the data supplied on an annual cycle, data from the Student Loans Account Manager (SLAM) (which was involved in student loan scheme account management up until 2000) is held permanently by Statistics New Zealand for the purposes of the annual reload of the integrated dataset. This historical data, while originating from SLAM, has come via two other agencies – Inland Revenue and the Ministry of Education. The SLAM data supplied by Inland Revenue for the years 1997 to 2000 includes:

- name, date of birth, education provider code, student ID, SLAM ID, Inland Revenue number, sex and ethnicity.

62. The SLAM data supplied by the Ministry of Education for the years 1992 to 1999 included a larger number of variables. Those items that are used in the production of the integrated dataset include:

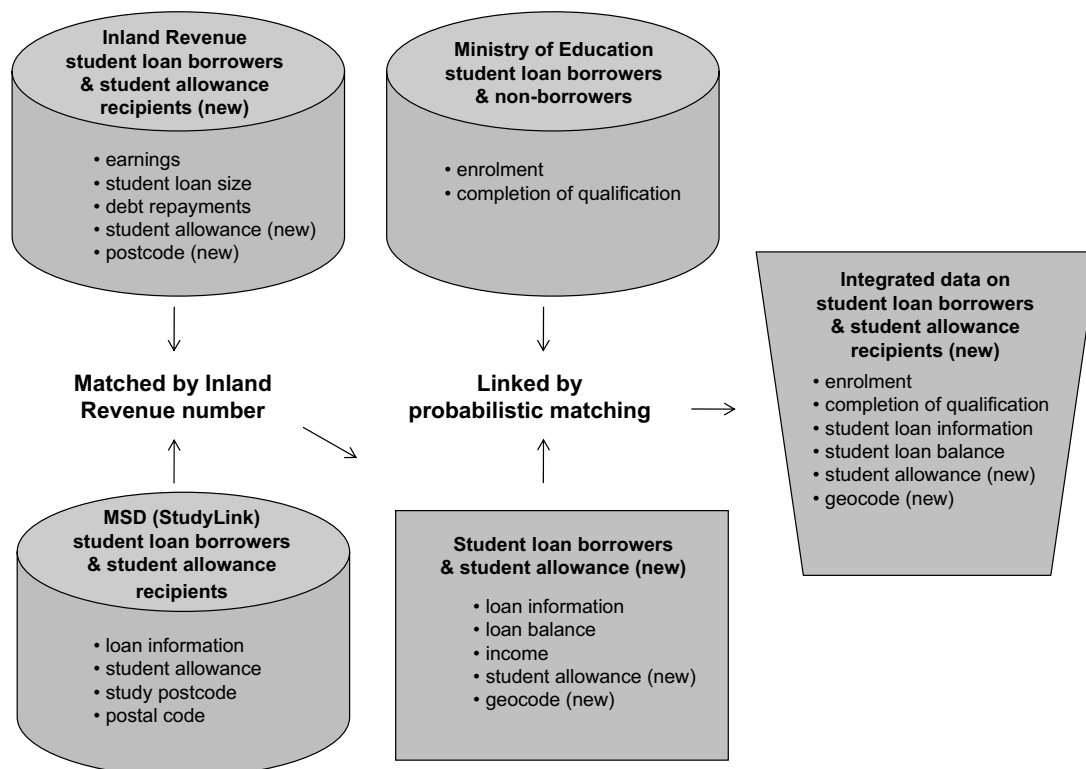
- date of birth, term address postcode, permanent address postcode, education provider code, student ID, Inland Revenue number, sex, ethnicity, nature of attendance, study start date, study end date, and study programme code
- annual fees entitlement, annual course-related costs entitlement, annual living costs entitlement, total annual drawings, annual course-related costs drawn, annual living costs drawn, annual interest accrued, annual repayments, and refunds.

63. In addition, the Ministry of Education-supplied SLAM files contained a large number of variables that have not proved useful either for linking or analysis. These data items have become redundant due to changes to the administration of the loans scheme, or because of methodological decisions concerning the best approach to producing the integrated dataset. In the main, this data comprises financial information in an alternative format. For example, annual entitlement and drawings totals have been used in the dataset rather than the monthly amounts that also appear in the files supplied. At one time this information would have been useful in assessing the feasibility of different possible approaches to integrating SLAM data.

64. All of the unit records containing SLAM data from both agencies include an ID number assigned by SLAM to identify individual borrowers. This is not used in the data integration and remains in the original file contained on the raw data server.

4.3 Data integration

65. The data integration process is illustrated in the following diagram.



66. Inland Revenue data is linked to StudyLink data (or SLAM data prior to 2000) using the Inland Revenue number. This is an exact matching procedure. The results are then matched to Ministry of Education data probabilistically, based on year of study/borrowing, student ID, education provider code, name and initials, date of birth, sex, ethnicity and, where available, Inland Revenue number (not all Ministry of Education records include an Inland Revenue number as it is not mandatory for students to provide this information). Probabilistic matching is used because the Ministry of Education data has no unique identifier in common with the Inland Revenue/StudyLink loans data. (The Inland Revenue number data collected by the Ministry of Education does not have sufficient coverage or quality for exact matching.) The probabilistic matching procedure allows records to be linked when some of the matching variables are not unique, have incorrect values, or are missing.

67. The link rate is the proportion of records that are successfully linked. During the latest annual reload the average link rate after both exact and probabilistic matching was

90 percent for the 1997 to 1999 period, and 97 percent for the 2000 to 2005 period. The difference in link rates is mostly due to the lack of education data for students at private training establishments prior to 2000. The overall link rate was 94 percent.

68. The integrated dataset contained data for some 705,000 people who have participated in the allowances or loan schemes. Approximately 6 percent had received an allowance but never borrowed through the loan scheme, and so would not have been included in the dataset if it had not been extended to incorporate student allowances data. The majority of student allowance recipients have also had a student loan and would have been included in the dataset regardless of this extension.

69. Statistics New Zealand staff have the sole responsibility for undertaking the data integration. As noted above, once this is complete, the integrated dataset resides on the clean data server. It does not include unique identifiers or name, date of birth or address, which are contained on the raw data server. But it does include the Statistics New Zealand assigned identifiers, and several variables derived from date of birth and address information: year of birth, month of birth, territorial authority, and region. The information in the integrated dataset, although still unit-record-level data, is therefore often referred to as anonymised data because it does not contain name and address information.

70. In some cases, it may still be possible for individuals to be identified because they have a combination of characteristics that is very unusual or unique. For example, if there was only one student over 75 studying a given subject at a particular institution in a certain year, then that unique set of characteristics may render the information about that person identifiable. To prevent such information being released, results are aggregated for publication and the data in the aggregated output tables is subject to confidentialisation procedures, including random rounding of population counts. Dollar amounts are also rounded and in some cases suppressed (that is, blanked out) to ensure confidentiality. Access to the anonymised unit-record data for research purposes is managed through Statistics New Zealand's data laboratory and microdata access protocols.

71. These measures also help to mitigate another problem that can occur. In probabilistic matching, it is possible for someone's loan data to be erroneously linked with the wrong education record. This is called a false positive match. If someone were to be recognised from data where this had happened, such information would be false and potentially misleading. Two methodological reports have been published on the Statistics New Zealand website describing and evaluating the data integration methodology in detail. The

false positive rate is also monitored during each annual reload and has been consistently estimated at less than 1 percent, meaning that overall the linked data is of high quality for statistical purposes.

4.4 Use and dissemination

72. The information obtained from supply agencies is used to produce the Integrated Dataset on Student Loans and Allowances. The dataset itself is used to produce aggregated statistical outputs and a copy is placed in the data laboratory shortly after the annual release. Access to this anonymised unit-record data is in accordance with Statistics New Zealand's microdata access protocols. (See appendix 2.) The raw data may also be used in the development and testing of changes to the applications used to create the integrated dataset, before these changes are implemented.

73. The integrated dataset is created for a statistical and research purpose. (It provides, in particular, improved information for strategic policy, financial analysis and reporting, in relation to the student allowances and loan schemes.) This is the sole purpose for which personal information is obtained on student loan and allowance recipients.

74. Some government departments use data matching exercises for administrative purposes, but this is not the case with the integrated dataset. The data will not be available for any administrative purposes, and there will be no transmission of information on identified individuals back to the supply agencies.

75. Access to identified unit-record data is restricted to a small team of Statistics New Zealand staff who process and integrate the data, and database administrators who maintain the technological infrastructure.

76. In keeping with policy, official statistics from the dataset will not be published in a form that would enable individuals to be identified. Results generated in the data laboratory are checked to ensure they meet a similar confidentiality standard before they can be removed.

4.4.1 Aggregated output tables

77. The release of statistical outputs from the integrated dataset follows the annual reload. An analyst initiates the process by running an application designed to generate specific tables for release. Before publication, the population counts are randomly rounded, and dollar amounts are rounded or in some cases suppressed (that is, blanked out), to remove any potentially identifiable information from the tables.

78. The release of statistical outputs is accompanied by a media release, commentary and technical notes, which give information on the background, integration methodology, extent and quality of data held by Statistics New Zealand. There have been five releases from the integrated dataset – in 2002, 2004, 2005, 2006, and on 3 April 2007.

79. Customised requests can be run on application. The release of customised tables is subject to the same confidentiality rules.

4.4.2 Data laboratory

80. There are four sections of the Statistics Act 1975 that make it possible to provide access to microdata for researchers. For practical purposes, only two of these are likely to apply to anonymised student loans and allowances unit-record data. They provide access to approved researchers employed by or contracted to government departments, including Statistics New Zealand.

81. **Government department researchers:** Access for government departments is at the discretion of the Government Statistician, and section 37C of the Statistics Act is considered to apply, meaning that access is “solely for bona fide research or statistical purposes pursuant to the functions and duties of that Government Department”. Employees and researchers who are contracted to a government department and who require access to microdata, must sign a declaration of secrecy. They are then considered to be employees of Statistics New Zealand for the purposes of implementing the security and confidentiality provisions of the Statistics Act.

82. **Researchers contracted to Statistics New Zealand:** Under section 21(3B) of the Statistics Act, an independent contractor under contract to Statistics New Zealand, and any employee of that contractor, can be required by the Government Statistician to make a statutory declaration of secrecy. They are then considered to be employees of Statistics New Zealand for the purposes of the security and confidentiality provisions of the Statistics Act. This provision enables Statistics New Zealand to provide researchers with access to microdata and requires them to protect the information in the same way as employees of Statistics New Zealand are obliged to do. Such access is granted to researchers undertaking work considered by Statistics New Zealand to fit with the purpose of Statistics New Zealand and which is of significant public interest.

83. There must be an intention to make the research available in the public domain for access to be granted. Requests for access to integrated datasets are also referred to the relevant steward agencies of the source datasets. Agreement is required from Inland Revenue, in particular, before access can be granted.

84. Datalab output is checked prior to removal to ensure it meets the relevant confidentiality requirements, principally, that it does not contain identifiable information. Draft publications containing the output must be sent to the datalab administrator and are checked for any changes or combining of outputs that result in confidentiality breaches.

4.5 Storage and retention

85. Under the Privacy Act 1993, Information Privacy Principle 5 requires that all reasonable steps be taken to ensure that personal information held by an agency is protected against loss; unauthorised access, use, modification or disclosure; or other misuse. Government departments are required to comply with the Security in the Government Sector (SIGS) protocols. A Statistics New Zealand security office functions to actively audit and review security processes and address new and emergent threats.

86. The source data held on the raw data server is retained for 15 months and destroyed when the following year's update is published.

87. An exception to this is the SLAM data that is held permanently on the raw data server. This is identified unit-record data on student loan borrowing for the 1992 to 1999/2000 period from the Student Loans Account Manager, which no longer exists. (Student loan scheme account management has been taken over by StudyLink.) Continuing to hold these datasets is currently the most practical solution to having the information on hand each year for the annual reload, given that as historical data it will never be updated, improved or changed. The original CDs have also been retained.

88. Regular backups are made to protect the data against power outages, hard drive failures and other IT problems. These are stored securely for a specified time, up to 12 months. The full database is also archived each year, and stored indefinitely.

4.5.1 Protection against loss

89. Within the student loans and allowances process, the CDs on which the original source data is delivered to Statistics New Zealand are retained securely until the data has been loaded and checked, and are then destroyed. CDs are hand-delivered to guard against loss. From 2007/08 the delivery will be by secure courier, and there will be a clear process to log the receipt and destruction of each CD, which will add an additional safeguard. Enhancements to the documentation process in June 2007 address this issue, and the function of receiving the data will be located in Statistics New Zealand's Integrated Data Collection area, which specialises in the role. There is currently no electronic transmission of unit-record data to Statistics New Zealand in this process, but the option of secure electronic transmission will be assessed in future. For example, the Government Shared

Network will link government agencies by a secure high-speed network and may offer an appropriate service for student loans and allowances data exchange. An initial set of services will be released in 2007.

4.5.2 Unauthorised access, use, modification or disclosure

90. Statistics New Zealand has in place well-established policies, procedures and systems to ensure adequate measures of physical and electronic security.

4.5.2.1 Unauthorised persons external to the organisation gaining access to data

91. All Statistics New Zealand buildings are subject to physical security systems that control entry to the premises, and to sections of premises, to authorised persons. All Statistics New Zealand employees are required to display Statistics New Zealand identification at all times. Visitors to Statistics New Zealand are subject to strict registration, and supervision procedures and systems that ensure their activities are confined to legitimate business. Access to Statistics New Zealand IT systems is password protected.

4.5.2.2 Unauthorised persons from within the organisation gaining access to data

92. The organisation operates a need-to-know policy that prohibits any staff, other than those actually working on a particular survey or process, from seeing confidential information. The system architecture has implemented role-based access control for all read and write activities related to the student loans and allowances data. Employees are trained and reminded to be security conscious, for example locking computer screens when away from their desks so that confidential information cannot be seen by others.

4.5.2.3 Authorised persons from within the organisation misusing their access to obtain information for personal interest or gain

93. Staff undertake security training and are trained on-the-job as confidentiality practitioners. Confidentiality courses are run frequently in-house and are available to staff, including awareness of security and privacy principles.

4.6 Summary

94. The significant privacy risks identified in this section are summarised below:

Privacy risk	Potential risk/impact	Risk mitigated by:	Assessment of residual risk
The risk of a breach of security or confidentiality involving personal information while in transit to or under the custodianship of Statistics New Zealand	Low/High	Standard Statistics New Zealand physical and data security practices augmented by: <ul style="list-style-type: none"> • data delivery by secure courier by 2007/08; documentation of procedures for receipt, storage and destruction of CDs completed by June 2007 • data being stored and used on servers dedicated to student loans and allowances • access granted to data on a need-to-know basis • integrated unit-record data is anonymised. Access to data laboratory strictly controlled • all releases checked to ensure confidentiality is preserved 	Low
Potential public concern about, or rejection of, the legitimacy of the use of this information (note: any adverse public reaction might not only impact on the student loans and allowances process, but could also conceivably impact on the reputation of the agencies involved and the ability of government agencies in general to link data from unrelated sources in the future)	Low/High	<ul style="list-style-type: none"> • adherence to Cabinet directive [CAB (97) M31/14] • compliance with Statistics Act requirements and Statistics New Zealand protocols • ongoing consultation with the Privacy Commissioner and data stewards • public openness regarding the dataset, with several public releases since 2002 • the removal of all identifying information from personal records prior to integration • access only granted to data on a need-to-know basis • processing requires a relatively small team • restrictions on data use and outputs of the process • data retention and destruction practices 	Low
Potential concern by the Office of the Privacy Commissioner regarding proposed procedures to: <ul style="list-style-type: none"> • replace individuals' Inland Revenue numbers with Statistics New Zealand-generated identifiers 	Low/Medium	<ul style="list-style-type: none"> • there is limited risk to individual privacy arising from this issue 	Low

5 Additional privacy-enhancing responses

95. Enhancement to the documentation of the student loans and allowances process in June 2007 include security procedures.

6 Future compliance mechanisms

96. All data integration projects must be approved by the Government Statistician. Statistics New Zealand's data integration policy requires that ongoing data integration work must be reviewed at least once every two years, to assess whether the benefits continue to outweigh the privacy concerns and risks.

97. A new business case must be approved before integrated data can be used for purposes not specified in the original business case.

7 Conclusion

98. The Integrated Dataset on Student Loans and Allowances is an ongoing data integration process that involves reloading, integrating, and releasing statistical information on an annual cycle. The integrated dataset is used for statistical and research purposes only. The benefits provided by the information on student loan borrowers, its extension to cover student allowance recipients and the supply of this information for some years beyond the point of leaving study, have been established and confirmed with stakeholders. In processing the data, the risk to privacy is minimised by restricting access to only those staff members who need to work on the data. Procedures are designed to remove identifying information and encrypt/replace unique identifiers from unit records before the data are integrated. High standards of security are maintained in Statistics New Zealand's physical and electronic environments. Strong confidentiality controls are applied to all outputs.

99. Risks to privacy are associated with: the security of data in Statistics New Zealand's custody and in transit, who may see the information, unauthorised access to or misuse of the data, and confidentiality.

100. These risks are managed by strong security and confidentiality processes within the organisation. Access to data is on a need-to-know basis, implemented through user authentication and role-based access controls for all read and write activities. All staff sign a declaration of secrecy, and receive security training. Training related to confidentiality is provided on-the-job and training courses are also available that incorporate awareness of

privacy principles. A security office monitors compliance with security policies, carries out audit functions and investigates actual and potential security incidents.

101. In some areas the following procedures could be strengthened:

- recording the receipt and destruction of CDs would add an additional safeguard against loss
- recording access and changes made to the database would provide a check for misuse.

102. The need for records of the receipt and destruction of CDs will be addressed through a documented procedure in future. The option for adding the functionality to record data access and changes onto the current system is technically available. An expert review of security risks and responses specific to the newly combined student loans and allowances database/process is also an option that could be explored.

Appendix 1: Data integration policy

Statistics New Zealand's data integration policy is available on the Statistics New Zealand website at: <http://www.stats.govt.nz/about-us/policies-and-guidelines/data-integration-policy/default.htm>

Appendix 2: Microdata access protocols

Statistics New Zealand's microdata access protocols are available on the Statistics New Zealand website at: <http://www.stats.govt.nz/about-us/policies-and-guidelines/general/microdata-access-protocols.htm>

Appendix 3: Privacy checklist

<ul style="list-style-type: none"> • Have security procedures for the collection, transmission, storage and disposal of personal information, and access to it, been documented? 	<ul style="list-style-type: none"> • Development of comprehensive documentation in June 2007.
<ul style="list-style-type: none"> • Are privacy controls in place for the project? 	<ul style="list-style-type: none"> • Yes, eg need-to-know policy and procedures for personal information access, physical security and access controls, IT security and access controls.
<ul style="list-style-type: none"> • Have technological tools and system design techniques that may enhance both privacy and security been considered? 	<ul style="list-style-type: none"> • Yes, eg encryption and replacement of unique identifiers, separate dedicated servers, restricted directories, technologies of anonymity.
<ul style="list-style-type: none"> • Has there been an expert review of all the security risks and the reasonableness of countermeasures to secure the system against unauthorised or improper collection, access, modification, use, disclosure or disposal? 	<ul style="list-style-type: none"> • Computer systems are premised on Statistics New Zealand's security standards, developed on an organisation-wide basis and monitored by the department's security office, which is headed by a senior consultant. An expert review has not been conducted specifically on the student loans and allowances system.
<ul style="list-style-type: none"> • Have staff been trained in requirements for protecting personal information and are they aware of policies regarding breaches of security or confidentiality? 	<ul style="list-style-type: none"> • Staff undergo security training, are trained on-the-job as confidentiality practitioners, and work with specialists within the department to ensure confidentiality is maintained in information that is to be released. Confidentiality courses are run frequently in-house and are available to staff.
<ul style="list-style-type: none"> • Are there authorisation controls defining which staff may add, change or delete information from records? 	<ul style="list-style-type: none"> • Yes.
<ul style="list-style-type: none"> • Is the system designed so that access and changes to data can be audited by date and user identification? Does the system 'footprint' inspection of records and provide an audit trail? 	<ul style="list-style-type: none"> • Not currently. It would be possible to build this functionality on.
<ul style="list-style-type: none"> • Are user accounts, access rights and security authorisations controlled and recorded by an accountable systems or records management process? 	<ul style="list-style-type: none"> • All rights must be approved by the project manager or data custodian, who can request from the system administrator a record of all rights currently in place.
<ul style="list-style-type: none"> • Are access rights only provided to users who actually require access for the stated purposes of collection or consistent purposes? Is user access to personal information limited to that required to discharge the assigned functions? 	<ul style="list-style-type: none"> • Yes.
<ul style="list-style-type: none"> • Are the security measures commensurate with the sensitivity of the information recorded? 	<ul style="list-style-type: none"> • The system provides an appropriately secure environment. Documentation of procedures for data arriving on CD will be available in the future.

Appendix 3: Privacy checklist (continued)

<ul style="list-style-type: none"> • Are there contingency plans and mechanisms in place to identify security breaches or disclosures of personal information in error? Are there mechanisms in place to notify security breaches to relevant parties to enable them to mitigate collateral risks? 	<ul style="list-style-type: none"> • Security is a core value within Statistics New Zealand and staff are required to be vigilant to any potential security breaches. • Statistics New Zealand's security office regularly monitors for compliance with security policies, carries out audit functions and investigates actual and potential security incidents. It has escalation procedures for the management and reporting of security incidents. • A memorandum of understanding between Statistics New Zealand and the data supply agencies specifies both technical and policy/relationship contacts within each department to consider any issues that may arise.
<ul style="list-style-type: none"> • Are there adequate ongoing resources budgeted for security upgrades with performance indicators in systems maintenance plans? 	<ul style="list-style-type: none"> • The documentation of procedures, including those around security, is the key step for the future and is resourced. Production of performance indicators is not anticipated at this time.
<ul style="list-style-type: none"> • What steps are to be taken to make the public aware of the project? Are individuals covered in the source datasets aware of the data use? 	<ul style="list-style-type: none"> • Data on student loan borrowers has been used in the project for several years, and several reports including information about the project have been published on Statistics New Zealand's website. • Key stakeholders have been consulted with regard to the inclusion of student allowances data. A statistical release including allowance data was published on Statistics New Zealand's website in April 2007, accompanied by a media release. • Inland Revenue forms state that Inland Revenue has an agreement to supply information to Statistics New Zealand for statistical purposes.

Glossary

Term	Definition
Anonymised unit-record data	Unit-record data with direct identifiers (such as name and address) removed
EFTS	Equivalent full-time student
IPP	Information privacy principle
ISCED	International Standard Classification of Education
Microdata	Unit-record and aggregate data that cannot be made publicly available, unless data is modified to protect respondent confidentiality
NSCED	An educational qualification classification code
NSN	National student number
NZSCED	New Zealand Standard Classification of Education
SAS	Statistical Analysis System. Software used for data management and production of statistics
SLAM	Student Loan Account Manager – prior to 2000, an agency (or agencies) contracted to manage student loan accounts
Suppression (of cells in a table)	Primary cell suppression protects sensitive cells by blanking them out
SWN	Social welfare number, assigned by the Ministry of Social Development
TFEA decile	Targeted Funding for Educational Achievement decile. A rating from 1 to 10 assigned to each school by the Ministry of Education for funding purposes
TSEC_SN student number	A Ministry of Education student number that has been created as an identifier – and has been assigned with reference to the NSN for 2003 and later years, that is, since the introduction of the NSN

References

Controller and Auditor-General (2000). *Student Loan Scheme – Publicly Available Accountability Information*.

Department of the Prime Minister and Cabinet (2002). *Security in the Government Sector*.

Ministry of Education: Education Counts

http://www.educationcounts.edcentre.govt.nz/_data/assets/file/0004/9553/about-the-integrated-dataset.doc (1 November 2007)

Ministry of Education, Inland Revenue, and Work and Income New Zealand (2000). *Student Loan Scheme Annual Report to 30 June 2000*.

Office of the Privacy Commissioner (2002). *Privacy Impact Assessment Handbook*.

Statistics New Zealand (2007). *Student Loans and Allowances: 2005*.

Statistics New Zealand (2006). *Student Loan Borrowers 1997–2004*.

Statistics New Zealand (2005). *Matching and Results of the Student Loans Data Integration Project 2000–2002*.

Statistics New Zealand (2005). *Matching and Results of the Student Loans Data Integration Project 1997–1999*.

Statistics New Zealand (2003). *Linked Employer-Employee Data Project: Privacy Impact Assessment*.