Working Paper No. 2
English only

UN/ECE Work Session on Statistical Metadata
(Washington, D.C., United States, 28-30 November 2000)

Topic (i): Statistical metadata for dissemination

Disseminating Data Quality Information via Metadata - ABS Experience

Submitted by the Australian Bureau of Statistics

1. INTRODUCTION

1. In common with most national statistical agencies, the ABS is witnessing continuing growth in the demand for statistical data. Not only are there more users seeking information across a wider range of topics, many of these users are also trying to extract more information out of existing statistical sources than ever before.

2. While this trend is gratifying for statisticians, generally, it raises concerns about the potential for clients to attempt to use these data for purposes beyond their data quality limits. Examples include requests for small area figures sought from collections which were only intended to support analysis in respect of higher level geographic regions; detailed commodity information from foreign trade statistics where editing and confidentiality constraints start to impact on their coverage and accuracy; and detailed industry breakdowns from derived statistical products such as input-output tables and national accounts figures.

3. At the same time, we need to move towards greater use of self-help facilities, dissemination via intermediaries, and a more cost effective, self-service, subscription system, as well as improving the efficiency of our information consultancy capability. It is therefore important that we build in procedures that alert users to significant data quality issues, and if necessary to prevent them from accessing doubtful datasets.

4. The ABS is well placed by way of publically accessible containers of data and metadata, with our ***AusStats*** and ***ABS@*** electronic products, while enhancements to our publication assembly and subscription service facilities have also been put into production recently. These facilities incorporate links between data and metadata.

5. This paper looks at the potential for metadata to draw users' attention to various data quality issues. In particular, it reports on ABS experience with efforts to attach metadata to our data objects which then flow through to our statistical outputs.

2. QUALIFYING AND QUANTIFYING STATISTICAL DATA QUALITY

6. Considerable discussion has been devoted to statistical quality. A major milestone in this discussion was the development of a data quality framework by Statistics Canada. This framework comprises six elements, viz:

- relevance
- accuracy

- timeliness
- accessibility
- interpretability
- coherence

7. The other key feature about data quality is that it is defined by the user. This is fine in a conceptual sense, but we are all aware that different users can and do have different requirements. For the statistical agency, this means that we have to exercise judgement to balance the needs for meeting our "duty of care" responsibilities, against the desire to present our data in a simple, convenient and readily accessible form.

8. Typically, we do this by providing more or less information with our data, depending on the attributes of the data and their users. For example, in Australia our quarterly wholesale price index for copper materials bulletin comprises four pages, containing one table of data, some main features, notes about the data, and explanatory notes about the collection. By contrast, a recent new release - Tourism Satellite Accounts - contains fifteen pages of tables and nearly sixty pages of accompanying text!

9. For the purposes of this paper, I propose to concentrate on the quality attributes of cells of data shown in our published tables. In particular, I want to focus on four potentially significant aspects:

i measures of sampling variability;
ii measures of coverage;
iii indicators of editing quality; and
iv impacts of confidentiality

Each of these aspects is described briefly, below.

10. Most agencies have a standard procedure for indicating the sampling variability of estimates derived from sample surveys. In Australia, we provide measures of estimated relative standard error. Often we annotate estimates which have a high but acceptable level of sampling error, and suppress those which have an excessive level. In other cases, we provide a table of "indicative" standard errors and leave it up to the user to decide whether the values of interest to them have an acceptable level of sampling variability.

11. For some collections, we need to constrain the coverage of contributing units based on perhaps a size criteria. For example, in our agricultural commodity survey, small establishments are excluded for the collection because their contribution, even in aggregate, is negligible to the key measures being counted. Nevertheless, for certain items (say, cauliflower production), the contribution of these excluded units can represent a significant proportion of the true total value of those items. We periodically attempt to estimate this by means of coverage (or undercoverage) surveys.

12. For similar reasons, we may also limit the amount of editing that takes place on certain items in a collection where the aggregate of that item accounts for a non-significant proportion of the total activity being measured. For example, commodities which have an export value of less than $A250,000 pa no longer have edits applied to the reported unit values in our merchandise exports collection. This could result in reported values being too high, low or volatile, as well as in possible errors in the reported State of origin or country of final destination, among other things. While the impact of such procedures is calculated to be negligible to the total value of Australian exports, certain data items at the lowest levels of disaggregation are likely to be more seriously affected.

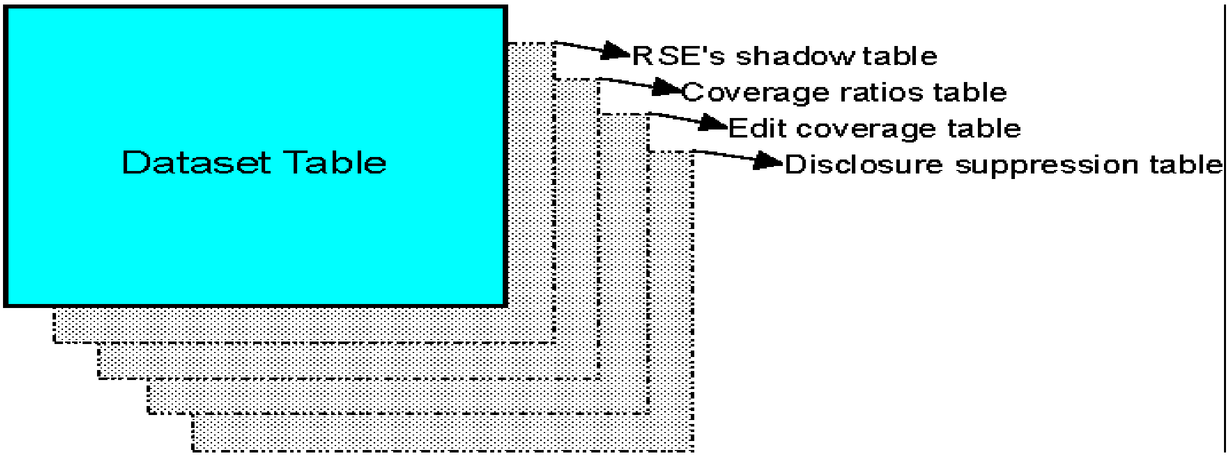
13. Finally, a number of our collections need to apply disclosure analysis and avoidance

procedures to identify and suppress identifiable information at detailed levels of disaggregation. When this occurs, individual cells may be suppressed or adjusted to avoid the disclosure of information that might identify the reported values for a particular contributor or contributors. This occurs in our merchandise trade collections, among others, which can impact on the accuracy of particular row and column totals in a table, such as exports of a commodity from a particular State.

14. Users need to be made aware when such limitations impact on the data that they are interested in.

15. Metadata provides a good way of handling these particular data quality impacts. Measures and indicators of these quality attributes can be supported by means of annotations down to the individual cell level in a table. That is, it is possible to attach multiple annotations to all cells in a table to reflect one or more of the various quality attributes.

16. Diagrammatically, this may be represented by a table containing the relevant values in a statistical dataset. Accompanying such a table is a set of "shadow tables" which hold the relevant measures and indicators for the respective data quality attributes.



17. The remainder of this paper discusses our progress to date on supporting some of these measures via metadata in our evolving dissemination systems.

3. HANDLING DATA QUALITY

18. Traditionally, statistical dissemination has been based on hard-copy publications. Such products enabled the statistical agency to provide, usually in a consolidated and standardised form, a range of descriptive, explanatory, technical and qualitative notes relating to the body of information being released. Agencies can thereby claim to have exercised a high degree of "duty of care" regarding the quality of information released.

19. However, issues such as the dearth of information about non-sampling errors, and dissemination of data by secondary providers may mean that end-users of these data were not always getting the full story. In addition, of course, different users react differently to our caveats and concerns.

20. The ABS response to these data quality issues is manifold:

- new standards and protocols for describing and measuring various attributes of data quality are being investigated and developed;
- quantitative data quality measures are being specified and incorporated into survey processing systems, for delivery to the central information repository, the

ABSDB;

- output production processes are being automated to simplify product creation for authors and to apply corporate standards;
- the new dissemination systems are being required to source their data from the central repository. These systems subsequently draw information from the ABSDB in the ***rich table object*** (RTO) format. Data delivered in this format have their associated metadata embedded with them. These metadata, in turn, have the potential to drive aspects of the content, format and layout of such products; and
- the ABS is attempting to play a leadership role among other sources of official statistics, by setting an example in terms of presentation of data quality attributes and encouraging other agencies to adopt similar practices.

A short description of each of these initiatives is given below.

21. A major project commissioned earlier this year is our "Qualifying Quality" initiative. This project is attempting to articulate a set of measures that should be compiled by collection areas. These measures are intended to address various quality attributes derived from the data quality framework developed by Statistics Canada. Work is progressing on identifying the suite of measures that ought to be assembled, noting that different measures may apply to different collections. Examples of the kind of quality measures being considered in Attachment A, below. The other important aspect of this project is to develop an education strategy that will assist people (both statistical compilers within the ABS and out external clients) to better appreciate the importance of information on quality and how to use it.

22. In parallel with this project are the major processing system development projects for household and economic collections (Household Survey Facilities, HSF, and Survey Processing Environment for Economic Data, SPEED). The work plans of both project teams include the provision of facilities to generate various data quality measures such as sampling errors and response, processing and edit rates.

23. In addition, the ABSDB is being upgraded to support the integrated loading, storage and delivery of these measures with the associated collections and datasets held on our Information Warehouse.

24. The first of the new generation of output production systems - the *Publication Production Workbench, PPW* - has been developed and put into production over the past year. This system largely automates the publication assembly process, saving author areas from having to deal with presentation issues, as well applying corporate publication standards. Another new system - the *Generic Special Subscription Service, GSSS* - enables particular collection areas to specify and deliver any number of customised outputs from datasets held on the ABSDB to meet the recurring requirements of clients. In this case, certain subsets of metadata are drawn from the ABSDB and other sources dynamically (rather than in RTO form) to accompany the statistical output generated by GSSS. Further information about GSSS, is provided, below.

25. Finally, our new Corporate Plan places a high priority on setting a strong example for other Australian suppliers of official statistics in relation to good statistical practice. Key strategies listed against our Corporate Plan objective for expanding and improving the "national statistical service" include:

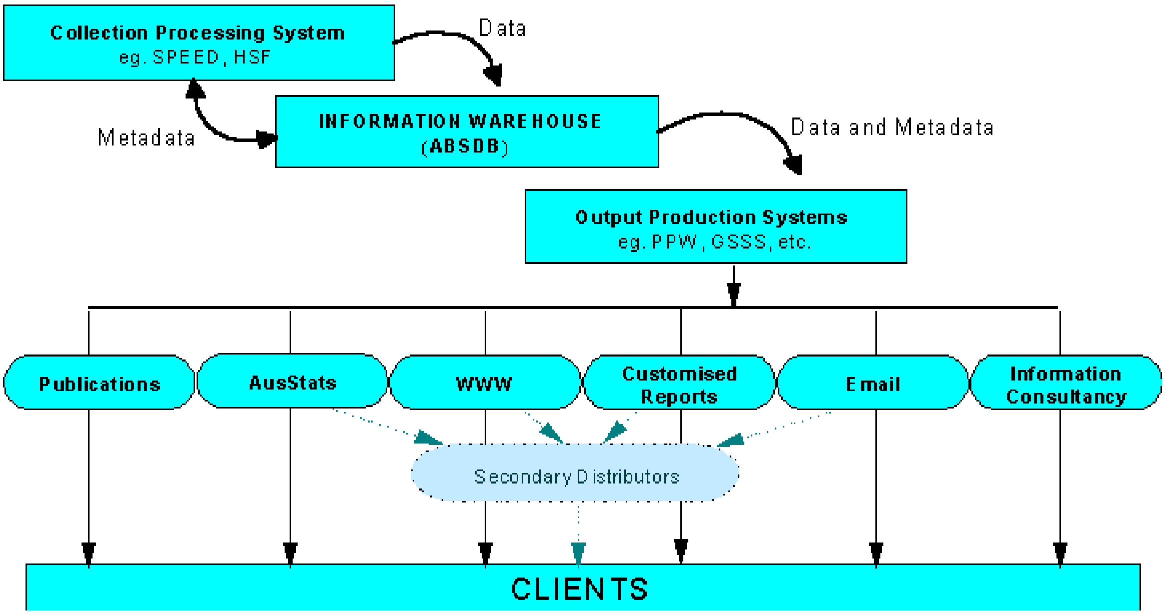
- developing and promoting a protocol which sets out the obligations of government agencies in the management and publication of statistical data from their administrative sources;

- promoting and supporting good statistical and data management practices;
- active participation in the development of standards for electronic information exchange and underlying data sets; and
- better utilisation of both public and private administrative and transactional data sources for national statistical service purposes.

4. ABS VISION FOR IMPLEMENTATION

26. The ideal scenario as far as ABS is concerned is that all collections should generate a comprehensive set of data quality measures that would be held in the central data and metadata repositories of the ABSDB. Thereafter, all statistical products would be created using standard corporate facilities, drawing the relevant data and metadata from the ABSDB. These products would present this information to clients according to agreed standards - depending on the dissemination medium (paper/electronic/etc) and the profile of user requirements for these products.

27. Diagrammatically, we might represent this vision as follows:



28. A key objective for the ABS that this diagram attempts to represent is the central importance of metadata in the statistical cycle. Thus, collection areas would both source much of their metadata from the ABSDB as well as update and increase their holdings of metadata in the Warehouse. Similarly, much of the metadata needed to assemble information reports and outputs would be drawn from the ABSDB, also.

29. In relation to data quality information, the ABSDB is intended to support the capture and storage of relevant metadata. The intention is to incorporate the business logic into our systems so that annotations to individual cells that are subject to high variability and/or confidentiality would be attached automatically where relevant. These annotations would then be delivered along with the statistical data into the various

dissemination products.

30. In reality, of course, only part of this vision has been achieved, although the development and deployment of the ABSDB, PPW and GSSS has advanced these goals a long way. The idea being pursued is to support a "pick-list" of data and metadata ***elements***; and a set of notional "templates" - one for each product - that we will incorporate in our output production processes. In some cases, subject matter areas may further select from a list of other components that can be included in respective products. For example, Sections or components of a publication might comprise: preface, abbreviations, introduction and background; analysis and summary of results; tables; conceptual framework; sources and methods; classification concordances; glossary and bibliography. Some components may be included in certain publications, but not in others.

31. The ABS is well placed to support this vision. Our Collection Management System (CMS) stores much of the contextual information for all ABS (and a large number of non-ABS) collections, from which relevant output "elements" and "components" can be drawn. Because of this, it will play an important role in gathering metadata relevant to quality assessment.

32. The remainder of this paper focuses on one of these systems - the Generic Special Subscription Service - to highlight a number of related issues.

5. THE ABS GENERIC SPECIAL SUBSCRIPTION SERVICE

33. The Generic Special Subscription Service (GSSS) is a system for managing subscriptions of specialised ABS reports. These reports are drawn from data and metadata held in the ABS Information Warehouse (ABSDB), and customised to meet the specific requirements of users for regular order. The idea behind this system is that it should provide a single environment, driven by an ABS information consultant, in which:

- client details and requirements details may be captured and held for both financial and output specification purposes;
- their specifications can be submitted to the output engine;
- associated information can be appended to this output;
- the relevant outputs can be packaged together into an integrated "whole" product for delivery to the client;
- this information is held in a secure repository, pending the passage of its embargo; and
- conversion of the information into the required format and medium, prior to its despatch to the client.

34. The system has been developed and deployed to support regular orders for international trade statistics and producer price index information. Other subject matter areas have recently started to inquire about using this system for their collections.

35. When a report is produced by GSSS for international trade statistics, tabulated data are accompanied by some relevant explanatory material; footnotes and annotations to explain where the client's report has been affected by confidentiality restrictions for the period covered (if applicable), or where they has been a category code change; the relevant sections of the commodity and other classifications shown in the table, and some code history information; and a list of glossary items.

36. A particular feature about this system is the way it dynamically holds and supports the production of these customised reports. GSSS combines the specified data with its descriptive metadata and some of the key data quality attributes of the data.

37. At a future stage, as well as supporting a wider range of collections, it is intended to draw more of the data quality metadata from the ABSDB, to support more extensive annotations and accompanying information.

6. CONCLUSION

38. Metadata provides a powerful way for statistical agencies to attach data quality information to their statistical outputs. It lends itself to supporting the six elements of the data quality framework. It also lends itself to customisation of the amount and detail of such information that we may want to attach to our different outputs.

39. The ABS has started to draw on its investment in the consolidation of data and metadata held in its Information Warehouse. Although only a limited amount of data quality metadata is currently shown in our main electronic products (*AusStats* and *ABS@*), our publication and subscription assembly systems are starting to draw both kinds of information from the ABSDB. Work is also underway on the development of a comprehensive list of data quality measures that the ABS will eventually want to draw from to meet its "duty of care" obligations.

40. This Work Session might like to consider whether there are opportunities for agencies to collaborate on the set of measures that ought to be compiled to support the data quality objective; the minimum standards that ought to be included in various delivery channels; and the templates that might be considered for different styles of dissemination.

Australian Bureau of Statistics
October 2000.

Attachment A

Quality Measures Currently Under Consideration

The following example has been provided as one possible option, recognising that they are not all encompassing and noting that it will always be important to be responsive to specific issues associated with each collection.

- Relevance - *Are these data measuring what I need?*
  - Scope
  - Frame issues
  - Concepts measured
  - Important classification issues
- Accuracy - *Are these data accurate enough for the purposes for which I need it?*
  - Level of sampling error (eg. Standard errors / relative standard errors / confidence intervals / statistical tests)
  - Relevant information on time series
  - Relevant non-sampling error
    - strata with high levels of non-response
  - Impacts of outliering
  - Size of revisions
- Timeliness - *Are the data I need available?*
  - Time after reference period
  - Schedule for revisions
- Coherence - *Are the data consistent over time or with other data sources?*
  - Level of consistency over time
  - Level of consistency with other related statistics
  - Quality measures available for comparisons
    - over time
    - with quality measures from other collections

APPENDIX 1 - Example of Quality Issue Summary for Monthly Labour Force Survey

Relevance

The Labour Force Survey (LFS) provides a measure of the currently economically active population (ie the labour force). This population is conceptually equivalent to the pool of labour available for the production of economic goods and services as defined for System of National Accounts (SNA) measures of economic output.

The Labour Force Survey collects information from most people aged 15 or more. It excludes people who are permanent defence force personnel, certain diplomats of overseas countries, overseas residents in Australia, and members of non-Australian defence forces and their dependants stationed in Australia.

All States and Territories in Australia are covered in the survey. The Cocos (Keeling) Islands and Christmas Islands, which are also part of Australia, are excluded. Similarly, Jervis Bay Territory has been excluded since July 1993, having previously been included in estimates for the Australian Capital Territory.

Accuracy

Level and movement estimates

Estimate	*Dec 1996 Level estimate*	*1 Standard error on level estimate*	*Relative standard error*	*2 Standard Error range on level estimate*	*Nov-Dec 96 Monthly movement*	*1 Standard error on monthly movement*	*2 Standard Error range on monthly movement*
Employed persons	8,401,500	±25,100	0.3%	8,351,300 8,451,700	+4,500	±19,600	-34,700 +43,700
Unemployed persons	794,500	±10,900	1.4%	772,700 816,300	+19,300	±8,100	+3,100 +35,500
Unemployment rate	8.6%	±0.1	..	8.4% 8.8%	+0.2 pts	±0.1	+0.0 pts +0.4 pts
Participation rate	63.6%	±0.2	..	63.2% 64.0%	+0.1 pts	±0.1	-0.1 pts +0.3 pts

Timeliness

Preliminary estimates from the Labour Force Survey are released in the first two weeks of the month following the survey, whereas non-preliminary estimates are released by the end of the calendar month following the survey. For instance, June statistics are available by the end of July.

Coherence

The Labour Force Survey has been conducted by the Australian Bureau of Statistics since November 1960. The survey has been national since February 1964 and has included the Indigenous population since August 1966. The survey was conducted on a quarterly basis until December 1977. In February 1978, the survey became monthly.

The Labour Force Survey questionnaire has been enhanced many times since 1960. The last enhancement occurred in April 1986 when the definition of employment was changed to include contributing family workers working 1-14 hours. A phase-in of telephone interviewing commenced in July 1996.

Labour Force Statistics are often compared to statistics from the Survey of Employment and Earnings and the Population Census. The Survey of Employment and Earnings estimates, from a sample of businesses, the number of jobs held by wage and salary earners and is thus not directly comparable to the Labour Force Survey which estimates, directly from a sample of people, the number of people in jobs. The broad concepts underlying the measures of the labour force and its components employment and unemployment are similar in the Population Census and Labour Force Survey. However, analysis of the 1991 Census estimates of unemployment have shown that labour force estimates derived from the Census differ to those from the Labour Force because they do not take into account the 'available to work' criterion.

APPENDIX 2 - Additional Data Quality Summaries

Example: APPENDIX 2 - DATA QUALITY ISSUES (PUBLICATION 3401.0)

NON-RESPONSE RATES PRIOR TO IMPUTATION AUGUST 1999(a)

OAD variables	*Incoming variables*	*Outgoing variables*
Citizenship (Nationality)	0.25	0.27
Country of birth	0.05	0.04
Age (Date of birth)	0.00	0.00
Sex	0.00	0.00
Marital Status(b)	32.60	43.57
Category of Travel	1.41	0.45
Permanent migrant		
Previous/future country of residence	0.00	0.00
State of intended address/lived	n.a.	n.a.
Overseas visitor		
Intended/actual length of stay	2.00	0.02
Main reason for journey	3.91	..
Country of residence	n.a.	..
State of intended address/in which most time was spent	n.a.	n.a.
Australian residents		
Actual/Intended time away from Australia	0.00	0.94
Main reason for journey	..	2.29
Country spent/intend to spend most time in	n.a.	n.a.
State of intended address/lived	n.a.	n.a.
Occupation(c)	0.00	0.00
Flight number or name of ship	0.00	0.00
Country of embarkation/disembarkation	0.60	0.19
Airport/Port of arrival/departure	0.00	0.00
Arrival/departure date	0.00	0.00
Whether intend to live in Australia for next 12 months	1.31	..

(a) Non-response rates are unweighted.
(b) Not available of Australia or New Zealand.
(c) Not available for short-term movements.

INTENDED LENGTH OF STAY / TIME AWAY FROM AUSTRALIA

Non-response rates are available for these data items from November 1998. For data prior to November 1998, imputation carried out as part of processing by the Department of Immigration and Multicultural Affairs (DIMA) has prevented reliable estimation of non-response rates for these two data items.

MAIN REASON FOR JOURNEY

Before the introduction of the redesigned passenger card in July 1998, 5% of short-term visitor arrivals, on average, were recorded as having a reason for journey of 'Other' or 'Not Stated'. This percentage rose to 14% for July, 16% in August and 29% in September 1998 as a result of processing problems. These problems have now been addressed by DIMA, with the percentage of 'Other' and 'Not Stated' dropping in October 1998 to 8% and 7% in November 1998. From the January 1999 issue of this publication, published figures (Table 3 and Table 9 in this publication) referencing these three months have been revised. The revised data were calculated by estimating the number of persons responding 'Other / Not Stated' using past trends for each country of citizenship and proportionally allocating any persons in excess of the estimated 'Other / Not Stated' total amongst the remaining categories. 'Not Stated' rates are now separately available from February 1999 onwards.

STATE IN WHICH MOST TIME WAS SPENT

For the months of August 1998, September 1998 and October 1998, data entry problems experienced by DIMA caused an overstatement of the Northern Territory as the main State of stay with a corresponding understatement for the remaining States and Territories. These numbers have returned in November 1998 to levels more comparable with previous years, with DIMA indicating that they have instigated data quality procedures to address this issue.

From the January 1999 issue of this publication, published figures (Table 8 in this publication) referencing these months have been revised. The revised data were calculated by estimating the number of persons indicating the Northern Territory as their main State of stay using past trends and proportionally allocating any persons in excess of these estimates amongst the remaining States and Territories.

LONG-TERM MIGRATION

Long-term migration for departing overseas visitors and arriving Australian residents has fallen markedly between 1997/98 and 1998/99. Investigation into the cause(s) is continuing, however it may be due to the more precise method of determining duration of stay using the new passenger cards and/or the Asian economic crisis.

SEPTEMBER 1998 PROCESSING

A problem was experienced in the processing of OAD data for movement dates between 6 September 1998 and 16 September 1998, following the introduction of changes to DIMA's input processing system. This problem may affect in the order of 10% of all September records used in estimation and result in incorrect details for citizenship, date of birth, sex and country of birth.

DATA IMPUTATIONS

Data are imputed for certain variables when no responses are recorded on the respective passenger cards. These variables and the information used to impute for them are listed in the table below.

DATA ITEM IMPUTATION

DATA ITEM	*IMPUTATION*
Category of travel	Includes references to citizenship (Australia, New Zealand, Other), corresponding migration visa, intended length of stay and whether intend to live in Australia for next 12 months
Intended length of stay	10 days
Country of residence	Country of departure, if it also matches country of citizenship
State lived / in which most time was spent	State of clearance