Australian Bureau of Statistics

1522.0

Information Paper

Quality Management of Statistical Outputs Produced From Administrative Data

Australia

March 2011

Information Paper

Quality Management of Statistical Outputs Produced From Administrative Data

Australia

March 2011

Brian Pink
Australian Statistician

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30AM (CANBERRA TIME) TUES 22 MAR 2011

ABS Catalogue No. 1522.0

Produced by the Australian Bureau of Statistics

INQUIRIES

- For further information about these and related statistics, contact the National Information and Referral Service on 1300 135 070.

CONTENTS

page

INTRODUCTION

PREAMBLE

The Australian Bureau of Statistics (ABS) is Australia's official statistical agency and is committed to leading a high quality statistical service to assist and encourage quality informed decision making. As part of this service the ABS is committed to providing leadership in quality management of statistical processes.

National statistical agencies need to find new and improved ways of acquiring information about their target populations to reduce respondent burden, maximise available data usage and find efficiencies in the use of taxpayers' money. This has led to the use of information that is collected for purposes other than that of a statistical nature. This type of information is often obtained from records or transactional data from government agencies, businesses or non-profit organisations which use the information for administrative purposes. Examples of administrative data include:

- registration of births, deaths and marriages;
- customs data (e.g. import and export records);
- taxation data;
- financial data (e.g. automatic teller machine transactions);
- social security benefit applications;
- school enrolments; and
- hospital records, to name a few.

This paper provides information on managing the quality of statistics produced from administrative data. It focusses on some principles and best practices to assist in the management of the acquisition of administrative data. It will be useful for any agency considering entering into a relationship regarding the supply and access of administrative data, as well as those agencies already involved in this area. A future publication is intended for release to cover processing issues which are not discussed in this paper.

This paper is a continuation of the ABS' role in providing guidance on frameworks and best practice in the use of statistics, including those produced from administrative data. It can be used in conjunction with other previously released ABS papers in the quality management series:

ABS Data Quality Framework, May 2009, (cat. no. 1520.0); and

Quality Management of Statistical Risk Using Quality Gates, Dec 2010, (cat. no. 1540.0).

Another useful source for quality management information is the National Statistical Services' *Data Quality Online* (http://www.nss.gov.au/dataquality/) tool which provides detailed information on the application of the *ABS Data Quality Framework*.

INTRODUCTION

Administrative data consists of information collected about a population either directly or indirectly. It is collected by organisations for different administrative purposes including:

INTRODUCTION *continued*

- for regulatory purposes (e.g. motor vehicle registration);
- for the administration of various government programs (e.g. social security benefits, education, and health);
- for legislative requirements (e.g. taxation declarations);
- for analysis (e.g. supermarket scanner data);
- as a by-product of transactions (e.g. electronic funds transfer at the point of sale (EFTPOS)).

The administrative records can also be used for statistical purposes and in doing so can offer important advantages over direct collection of data from the population concerned.

Some examples of data sourced from administrative systems which have been used in ABS publications include:

- Register of Births, Deaths and Marriages provides the ABS with data which is used in population estimates and projections. Data from this source can also be used to measure likely demand for services associated with population growth, such as housing, child-care and schools *(Births, Australia - ABS* (cat. no. 3301.0); *Deaths, Australia - ABS* (cat. no. 3302.0))
- Motor Vehicle Registries provide snapshots of their registers to be used as a population frame for the Survey of Motor Vehicle Use, (the data from the snapshot is published in *Motor Vehicle Census - ABS* (cat. no. 9309.0))
- Department of Education collects information on school enrolments (*Schools, Australia - ABS* (cat. no. 4220.0))
- Customs documentation completed for goods sent or received at an international sea or air port, provide the basis for overseas trade data and export and import statistics (*International Trade in Goods and Services* (cat. no. 5368.0))
- Local government authorities' collection of information on building approvals is used to construct an indicator of economic activity (*Building Approvals, Australia - ABS* (cat. no. 8731.0))
- The Australian Taxation Office collects information on personal and business taxation which is used by the ABS. Publications such as *Wage and Salary Earner Statistics for Small Areas, Australia* (cat. no. 5673.0.55.003), the *National Regional Profile* (cat. no. 1379.0.55.001) and *Estimates of Personal Income for Small Areas, Australia* (cat. no. 6524.0.55.002) are examples of some of the uses of this administrative data.

In many ways the use of administrative data for the production of statistics or statistical analysis is akin to the production of statistics from data collected directly from respondents, as is done with a census or survey. Many of the principles and methods of best practice for the various processes involved apply regardless of the source, and the approaches for managing quality are the same. However, there are important differences which include (but are not limited to) the addition of at least one extra party in the acquisition of data and the fact that the administrative data are not collected for the sole purpose of statistical analysis. These differences and others need to be taken into consideration when using administrative data.

This paper will define the major uses of administrative data for statistical purposes and the quality management techniques which are important in the production of statistics. While this paper focuses primarily on acquiring electronic administrative data the

INTRODUCTION *continued*

principles can be applied to other forms of administrative data such as hospital record forms. Later releases will examine the processing of administrative data throughout the whole statistical cycle.

The term "data custodian" will be used to represent those agencies or organisations who collect administrative records as part of their normal business process requirements. "Receiving agency" will refer to those agencies or organisations who utilise administrative data for statistical purposes. Receiving agencies may include national statistical agencies or other government agencies or businesses who use administrative data to produce statistical outputs. In some cases it is possible that the data custodian and the receiving agency are the same entity.

CHAPTER 1 USES OF ADMINISTRATIVE DATA

USES OF ADMINISTRATIVE DATA

As previously mentioned the ABS uses administrative data in quite a number of statistical publications. Other uses of administrative data aside from producing statistics directly from the collection include:

- Data validation;
- Editing;
- Imputation and substitution;
- Reducing respondent burden;
- Register maintenance and frame creation (e.g. Stratification variables);
- Weighting and estimation (e.g. data substitution, weighting / calibration, synthetic estimates); and
- The ability to show longitudinal effects.

Data validation

Administrative data can be used to help validate existing data sources at an aggregate or unit level. Validation can take the form of checking to see whether an estimate achieved from one collection seems reasonable in comparison with an already existing collection reporting either the same or a similar estimate.

Editing

Editing is the activity aimed at detecting, resolving and treating anomalies in data to help make the data ‘fit for purpose' (ABS 2009c). Once again, administrative data may be used to provide the answers to questions raised about anomalies in a particular data source, effectively replacing the original data or providing guidance as to the treatment of the data being investigated.

Imputation and substitution

Administrative data can be used for imputation purposes. Imputation is the process used to determine and assign replacement values for missing, invalid or inconsistent data (SDMX 2009). For example, where a survey has data that needs to be imputed, an administrative data source could be used to provide the information required.

Data substitution occurs when data for the same unit is used from a different source. For instance, a survey may ask a reporting unit a question and when the answers are being validated it may appear that the response to the survey is not plausible given the information known about that reporting unit from an administrative source. The administrative data source response may be substituted into the survey if it is considered more accurate.

USES OF ADMINISTRATIVE DATA *continued*

Data substitution also includes the use of administrative data to replace a subgroup of a population. This may occur where the administrative data are of sufficient quality and the data available adequately match the concepts being measured by the direct survey collection. This in turn creates efficiencies by reducing the size of the survey sample required and also reduces respondent burden.

Reducing respondent burden

"Response burden is an unavoidable part of survey research, but efforts to limit it can help maximise response rates. Shorter questionnaires can improve response rates, particularly if interviewers inform respondents that the interview will be short" (Public Works and Government Services Canada 2007). When an agency conducts a survey the respondents are often overwhelmed by the extensive questions asked of them, as well as the number and frequency of the surveys they have to participate in. Statistical agencies generally need this level of detail in order to meet the requirements of key users of the statistical output.

As mentioned data substitution is one way of reducing respondent burden. Another is through the linking of administrative data about the target population in conjunction with other data sources, be they administrative or survey, to create richer data sets. This means that more detailed information is available about the population without undertaking a new or improved survey to obtain the already existing information. This reduces the number of questions which are required to be completed by respondents and hence reduces respondent burden.

Register maintenance and frame creation

The creation and maintenance of registers and creation of frames (for sample selections) can be achieved through the use of administrative data. A frame can be a list, map or other specification of the units which define a population to be completely enumerated or sampled (SDMX 2009). The ABS uses some Australian Taxation Office data to maintain a register of all the businesses in Australia from which frames and samples for various business surveys are then created.

The creation of a frame using administrative data also allows the use of auxiliary information for purposes such as stratification or other sampling purposes if there is enough detailed information from the administrative source. The ABS uses information such as industry, state, sector (public or private), the number of payees in a business (number of employees) and turnover to help stratify the business frame in order to collect information from a representative sample of businesses.

Using administrative data in the creation of a frame and its stratification can assist in improving sampling efficiency for output purposes. This is because more information is known about the population and therefore enables the identification of units to be included or excluded from the sample more easily. It also improves the ability to produce small area statistics (e.g. geographical, particular industries etc.), as well as enabling the prioritising of reporting units (e.g. prioritisation of non-response follow up by size of business).

USES OF ADMINISTRATIVE DATA *continued*

Weighting and estimation

Using the administrative data as a frame allows benchmarks to be created for use in weighting / calibration estimates. This allows a sample taken from a population frame to be weighted to represent the population without having to enumerate the entire population. This can improve sampling and estimation efficiency by providing estimates of greater precision or reducing the sample size required.

The ability to show longitudinal effects

If the data custodians of the administrative data continue to collect the information about their population(s) of interest then the data can provide receiving agencies (including researchers) with the ability to analyse structural or situational changes over time for a specific population.

CHAPTER 2 COMMON ISSUES WITH ACQUIRING ADMINISTRATIVE DATA

COMMON ISSUES WITH ACQUIRING ADMINISTRATIVE DATA

Using administrative data in a statistical process can offer strategic and some statistically important advantages over direct collection of data from the population concerned. However, there are a number of issues associated with the use of administrative data in the production of statistical outputs. Some of the common issues with the use of administrative data include:

- Legislation;
- Availability of statistical information;
- Stability of data;
- Collection issues;
- Influencing the collection;
- Metadata availability;
- Systems Issues;
- Timeliness;
- Political influence;
- Large volume of records;
- Editing and transformation;
- Linking of records; and
- Confidentiality and Consent.

LEGISLATION

Administrative data that are collected by data custodians in different jurisdictions may have different legislative requirements associated with their use and acquisition. Each data custodian may have a different requirement for reporting on the topic of interest and also on the level of detail of data available to an external party. These differences in reporting and legislation need to be managed up front when considering acquiring administrative data. Restrictions placed on access and use of the administrative data by legislation must be understood before using administrative data.

Legislative issues to consider are the ability to follow up with a unit of the target population to an answer on their administration form, either by the receiving agency or the data custodian who manages the administrative records. Another issue to consider is whether access by the receiving agency to micro data to support further research and analysis if required is possible.

Negotiation needs to take place with the data custodians over these potential requirements as there are likely to be some issues to be addressed in order to undertake these types of investigations. These issues could include legislative or policy barriers, sensitivities about confidentiality and privacy, and lack of interest in supporting queries from receiving agencies by the data custodians.

LEGISLATION *continued*

Maintaining the confidentiality of individuals and businesses the data are collected about in any outputs produced from administrative data is extremely important. Once again, it is important to be aware of the legislative requirements of each jurisdiction and to ensure that processes are in place to meet them.

Keeping abreast of any changes in legislation which may impact on the administrative data source is important as these changes may impact on the data. For example a change in the requirements for receiving a benefit may lead to different behavioural patterns by applicants. Applicants may be more eager to complete administrative forms in a timely manner if this enables them to receive potential financial benefits quicker.

The extent to which an administrative organisation has the authority to collect information impacts on the quality of the data. A data custodian which has legal authority to collect information from respondents will have more complete data available for use by a receiving agency.

AVAILABILITY OF STATISTICAL INFORMATION

The administrative data source may not represent all of the population required by the receiving agency. Alternative methods and/or data sources may need to be used in conjunction with the administrative data in order to obtain the information required. It is also important to know at what level the data are being collected. For example if data are collected based on a post code this limits how the data can be grouped to represent the populations in different geographical areas that are not aligned with post code boundaries.

It is also important to be aware that the scope defined by the administrative data source may not be the same as the scope that is actually achieved by that data source due to issues with implementation and access to reporting units. Data custodians may not have a vested interest in following up where data are not recorded for some of the population because their focus is not on the use of the data for statistical purposes.

STABILITY OF DATA

When considering acquiring administrative data it is worth considering the history of stability in the data and the drivers of instability before deciding to proceed with its use. For example, the frequency with which the regulatory environment around the system is subject to change is an important consideration. This is because a receiving agency may invest a lot of resources into its ability to acquire the data only to discover that the data are not available a few years after the initial acquisition.

How often the questions, data items, quality of items and scope and coverage change within the administrative data impacts on a receiving agency's ability to maintain a time series and also longer term plans for data.

These changes can also impact on a receiving agency's reputation. An example of where the stability of administrative data have impacted on an organisation's reputation occurred when the Office for National Statistics in the United Kingdom attempted to use health records to measure change in real outputs of services. Continuous changes in the underlying coverage and definitions of the administrative processes generated statistics that had little credibility and eroded public confidence (ONS 2005).

STABILITY OF DATA *continued*

Other examples where business processes impact on the quality of the administrative data are when changes are made to the way the data are collected. For example, there have been changes in the legislation surrounding the exportation of goods from Australia. These changes have focussed on an increase in the security measures around the exportation of goods from Australia and have been in place since the early 2000's. As a result of these changes to the underlying business processes, the export data have become more detailed and accurate. Failure to explain the effects of these changes within subsequent statistical outputs could result in users misinterpreting any increase or decrease in export figures.

Whether the data custodian of the data will continue to collect and supply the administrative data for a long period of time is an issue for consideration as changes in the supplier of the administrative data will require relationships and requirements for the data to be reestablished.

COLLECTION ISSUES

Due to administrative data not being collected for the sole purpose of statistical analysis, issues arise in the collection of the data that need to be considered in order to manage its quality. In some instances questions about the same idea or topic used on administrative forms may differ across jurisdictions. In these instances it is important to understand these differences and how they impact on the reporting of the information.

Testing of questions, peoples' understanding of the concepts and quality of the form design may not have been undertaken in the case of administrative data collections. Hence, the reported information may not contribute to valid measures of the desired concepts.

Interviewer or administrator biases and errors may occur depending on who supplies the information which is captured. In some cases a third party may be filling in the data and may or may not seek answers from the unit of interest.

Where the target population are required to complete a form or otherwise respond to data items, the administrative data may suffer from partial non-response. This is due to missing fields where respondents either haven't understood the question, accidentally missed the question or refused to answer the question. In these cases it is hard to know the reason behind the non-response, which may impact on the ability to analyse and use the data. An output editing strategy may need to be developed in advance of acquiring the data in order to have procedures in place for how to deal with such issues. A discussion with the data custodian up front about whether they can contact the reporting units and obtain data for these missing values through follow up procedures is an essential discussion to undertake. This will enable a decision to be made as to how potential missing data will be managed.

Administrative data can also be collected as either flow or stock data. Flow data are accumulated during the reference period. For example birth data is a flow series as the number of births are tallied throughout the year to provide the accumulated total of births for the year. Stock data are measured at the end of a reference period. For example company net profits as at the end of the financial year (30 June). Knowing what type of data are being acquired is important from a data quality management perspective (e.g. duplicate records).

INFLUENCING THE COLLECTION

There is usually limited opportunity to influence the data by influencing the business processes underlying the creation of the administrative data. This might include things such as the size and diversity of the population for which the administrative data exists, the methods used for data collection (e.g. whether hospital records are self report or reported on behalf of someone else) and what information is gathered about an individual or business. Limited opportunity to influence the business processes underlying the creation of administrative data can have a significant impact on the quality of the resulting statistics produced, particularly their relevance, accuracy and coherence.

The decision to use administrative data should rest on an assessment of whether the noise generated in the data by the business processes that created them are of an acceptable level in relation to what the data are supposed to represent. 'Noise' generated refers to the measurement errors which have an impact on the quality of data by distancing the data from what they actually represent. For example, 'noise' may happen because of a transcription error that is not fixed prior to data being used, so an answer of twenty three point five (23.5) is recorded as thirty two point five (32.5).

There is also the cost benefit trade-off of spending money and effort to influence and change the up-stream process, e.g. convincing organisations to provide electronic data files. This needs to be considered in regards to the efficiencies to be gained and the ability to trust that the agreed processes have been undertaken.

Influencing data custodians can be a very difficult task and they may want compensation for any changes to systems that are designed to capture data in a more efficient way for the receiving agency.

METADATA AVAILABILITY

Metadata relating to what information is collected, and how it is transformed and processed to produce statistical output is critical for users' understanding the quality of the outputs produced. Best practice suggests that metadata should not be created retrospectively but rather created as a by product of various processes and stored internally. Critical information on concepts, sources and methods and quality issues should be made available along with the outputs produced. Directions to the availability of more detailed information for those users needing it should be included. While this approach applies to all statistical outputs regardless of their data source, there are usually some particular characteristics of outputs produced from administrative data which need to be drawn to the attention of users of the outputs. Any adjustments made to the data as a result of these features during the processing of the data should also be declared (e.g. editing, imputation etc.).

If possible it is important to obtain any information from the data custodians which can assist in interpreting the information accurately. It is also important to continually check that the reported structure of the file and its contents have not altered from one time period to the next.

SYSTEM ISSUES

Data custodians of administrative data are responsible for the management of their systems. This can impact on the receiving agency acquiring the data because changes made to the systems that collect and store the administrative data can impact negatively on a receiving agency's ability to extract and use the data for their purposes.

SYSTEM ISSUES *continued*

Effort by the receiving agency in acquiring the data in a format that is acceptable to their systems can be greatly reduced by encouraging and helping the data custodians to use standard data items and/or apply standard classifications and coding to the information. This is on the proviso that coded information is acquired and coded to a satisfactory standard.

Understanding the impact on quality and the cost involved when setting up the systems for processing administrative data is a very important step in any acquisition of administrative data. This is because if system issues are not considered in advance the benefit of obtaining administrative data may fail to be realised and in reality become more expensive than directly collecting the information from reporting units.

When initially seeking administrative data consideration should be given to the maturity of the system that is being used to collect the data and any potential issues in terms of the reporting ability of the data. For example whilst Electronic Funds Transfer at the Point Of Sale (EFTPOS) data has reached a stable level of market penetration, this was not the case over the first ten years when progressive uptake would have masked any change in real patterns of retail spending.

TIMELINESS

Administrative data may not be current. Often there is a delay between when the data are collected and when they are available for use by another party. The other issues in terms of timeliness of administrative data are the provision of the information by reporting units and the collection method. For example, the provision of administrative data may be subject to delays because of when reporting units of the target population submit or register the information. For example, tax data may require reporting to be undertaken quarterly. However, in some instances a taxpayer (i.e. reporting unit) may ask for an extension to this deadline. Hence, not all administrative data for the time period for which data are being received will be included in the data file acquired by the receiving agency. The extension data records may become available in the next quarter and will automatically be included in the data file at that time even though they correspond to an earlier time period.

Another issue that needs to be managed is that of duplicate records. As mentioned above, records may be included for a specific reference period at a later date. However, in some cases incomplete records are provided in one extraction of administrative data to a receiving agency. Hence, when the next extraction of administrative data occurs and those formerly incomplete records have been updated with the latest information the receiving agency ends up with two different records for the same reporting unit. This is how duplicate records can be created. These duplicate records need to be consolidated and the surplus record removed.

A decision needs to be made as to when the administrative data will be extracted for use by a receiving agency. As with any data there is a trade off between accuracy and timeliness. The longer a receiving agency waits to acquire the administrative data the more likely it is that the administrative data are more complete and accurate. However, this in turn delays the use of the administrative data for statistical purposes. As mentioned above, delays in the provision of complete data will always be apparent with administrative data. The important thing is to establish a set of guidelines for how to deal with the likely issue.

POLITICAL ISSUES

When obtaining and releasing administrative data it is important to be aware of the political sensitivities that may impact on its provision. This is particularly the case if the administrative data could be perceived in an unflattering way. For example, it may be especially sensitive if the data are used to compare one jurisdiction to another. It is important to consider how the data are going to be displayed and reported upon. Data custodians may want to review the way in which the data are reported and used publicly.

It is important to determine roles, ownership of outputs and governance of the process early on in the discussion of the acquisition of administrative data. This is because a data custodian may be quite sensitive about what the data are going to reveal and may want to have an influence on how the data are reported. In order to alleviate any potential reporting or general maintenance issues in the future, it is recommended that discussions over ownership and governance arrangements be established from the beginning.

Another issue that may arise due to political influences is the use of different definitions for reporting. Data custodians may use non-standard definitions for collecting information if there is a known issue with a particular concept or topic which might prove to be unflattering to the data custodian.

The independence and transparency of the data custodian in collecting the data will impact on the quality of that which is provided as well.

LARGE VOLUME OF RECORDS

Administrative data sets are often a very large source of information. A receiving agency will need to take into consideration the impact of a large amount of data being acquired and how the volume of the electronic data will be able to be handled by the existing computer systems. It is important to make sure that information is not lost in the transfer from one system to another.

It is also worth noting that the larger the data set, the lesser the ability of the receiving agency to ensure the validity of all of the records. Measures will need to be put in place which allow the agency to have an acceptable level of comfort in the validity of the administrative records without the receiving agency having to individually assess each record.

EDITING AND TRANSFORMATION

Administrative data may require editing to be undertaken before it can be utilised. Editing could take the form of fixing typographical errors, imputing missing values, dealing with outliers etc. These strategies may be undertaken either by the data custodian or by the receiving agency. In general, it can be quite time consuming and costly to edit administrative data especially if access to the original information is not possible. It is recommended that an editing strategy be developed and agreed to with the data custodian.

Transformation of administrative data is another issue that needs to be considered. In particular, the compatibility of the administrative data file with the receiving agency's corporate systems. It is recommended that negotiations be held between the data custodian and receiving agency to determine the requirements for delivery of the administrative data. These could range from specifying the file format, e.g. spreadsheet, to the order in which the data is presented within the file.

LINKING OF RECORDS

If a receiving agency wishes to link data from different sources, including administrative sources, investigations into the legal practicality of undertaking such a process will need to occur. If all legislative requirements are met and data linking can occur then rules for dealing with inconsistencies across records for the same entity need to be developed. Rules for dealing with records not exactly matched on the basis of unique identifiers also need to be developed in order to be able to use various pieces of information together. One of the difficulties with administrative data is the lack of consistency and coherence with other sources which is needed in order to undertake data linking.

CONFIDENTIALITY AND CONSENT

It is important when obtaining administrative data for statistical use that the information provided remains confidential and private no matter how the data ends up being utilised. This means that no business or individual should be able to be identified from any of the statistical releases based on the administrative data.

The provision of accurate and timely administrative data is often reliant on public goodwill and cooperation. The reporting units (e.g. persons, businesses) about whom the data are being collected should be advised about the intended use of the data that they are providing.

Concerns about the protection of confidential information can cause a decrease in public confidence and a subsequent loss of accurate data due to non-response. For example, perceived misuse of EFTPOS data will not only be bad for the reputation and later effectiveness of the receiving agency in their survey activities, but also lose customers for the data custodian supplying the EFTPOS data. This in turn would reduce the data custodian's willingness to respond to any future data requests.

Problems in regards to confidentiality can occur during the hand over of administrative data if appropriate measure aren't taken with the security of the data transfer. It is important to ensure that the systems used in the transfer of data are compatible to avoid excess data handling. It is also important to make sure that all privacy and other applicable legislation have been investigated and adhered to prior to the agreement of data acquisition.

When using administrative data to link existing data sources along with complying with all relevant legislation it is worth while considering and resolving the governance of the use and access of the newly linked data source to ensure that privacy is maintained.

Some methods for minimising privacy concerns and protecting confidentiality when using administrative data include:

- informing respondents of how their data will be used;
- following privacy principles and legislation (both of the receiving agency and data custodian);
- be sensitive to cultural differences;
- honour guarantees made to respondents;
- restrict access to unit record data (i.e. data relating to individual units); and
- develop guidelines on access and release of the data to ensure that no business or individual can be identified from any statistical release.

CHAPTER 3 MANAGEMENT OF DATA SOURCES USING THE ABS DATA QUALITY FRAMEWORK

DEFINE THE DATA NEED

In order to achieve a particular outcome a data need (requirement) needs to be defined in order to make sure that the outcomes are met. When considering using administrative data for statistical purposes it is important to define what the data needs are before assessing potential administrative data collections to fulfil this requirement.

The *ABS Data Quality Framework* (cat. no.1520.0) can be useful in assisting the creation of a list of questions to establish a data need in line with a given objective for obtaining the data in the first place.

To describe a data need it is useful to create a list of questions beneath each of the seven dimensions contained in the *ABS Data Quality Framework* (Institutional Environment, Relevance, Timeliness, Accuracy, Coherence, Interpretability, Accessibility). This helps to ensure that all aspects of quality are considered when defining a data need. These questions can then be used to help assess potential administrative data sources for their fitness for purpose for the required objectives.

Some examples of questions that may be beneficial to consider before acquiring administrative data are:

- *Institutional Environment*
  - From what type of organisation is it acceptable to obtain data?
  - Do we care how the data has been collected? i.e. direct collection or collected from a third party, face-to-face or online form.
- *Relevance*
  - What population are we interested in?
  - What geographic level of detail do we require?
  - What concepts and data items do we need?
- *Timeliness*
  - When do we need the data?
  - How recent do the data need to be in order to still be useful?
  - Will we want to get a time series of this information?
- *Accuracy*
  - Do we have an acceptable level of quality that is required for the data we require?
  - Do we want to include data that has been imputed?
- *Coherence*
  - Do we want to compare or combine several sources of data?
  - Do we need data that conforms to particular standards?
- *Interpretability*
  - Do we want a copy of the questions asked?
  - Do we require a list of definitions for the concepts measured?
- *Accessibility*

DEFINE THE DATA NEED *continued*

- What level of detail do we need for this data?
- What is our budget for purchasing data?
- What processes do we expect around the acquisition of the data? e.g. secure electronic transfer

For more examples of questions that could be considered when defining a data need beneath each dimension of the *ABS Data Quality Framework* please see '*Defining A Data Need*', on the National Statistical Services website (http://www.nss.gov.au/dataquality/PDFs/DQO_Needs.pdf).

ASSESSING A DATA SOURCE

When assessing a potential data source it is important to understand and document the aspects of the source which are critical to the design of the system which will be used to process the acquired data. It is also important to provide information which assists sound use of the data by helping users understand their quality and limitations. This includes the information likely to be extracted for statistical analysis and those features of the environment in which the administrative system exists which can influence quality e.g. the way the information is collected, public sensitivities about the collector or the system.

Similar to those questions that would be used for helping to clarify a data need, a list of questions for assessing a data source can also be compiled using the *ABS Data Quality Framework*.

An example of some questions for assessing a data source include, but are not limited to the following:

- *Institutional Environment*
  - What type of organisation collected the data?
  - Under what authority/legislation was the data collected?
  - Does the data get compiled by a different organisation from the data custodian that collected it? If so, what type of organisation is it?
  - Can confidentiality be maintained?
  - How was the data collected? e.g. self-report or proxy or automatic/observation.
- *Relevance*
  - Population – does it match user requirements? Is it close enough?
  - Do the data need transforming to match the required population e.g. aggregation to different units, or need adjustments or supplementary collection to cover the missing population? i.e. decide on scope and coverage and determine which records are in and which out.
  - Will longitudinal analysis be of value for meeting user needs? If so, some form of unique identifier will need to be used for each record to allow information to be built up over time.
  - Who or what was the data collected about?
  - Do the data meet the demographic requirements for the statistical purpose (i.e. age, sex, employment status, geographic locations, etc.)?
  - What was the original purpose for collecting the data?
  - Have standard classifications been used?
  - Are all records (census) required or will a sample be adequate?
- *Timeliness*
  - How often are the data collected?

ASSESSING A DATA SOURCE *continued*

- When are the data available?
- Over what period are the data collected?
- Are update or revisions to the data likely after release?
- What cut off dates should apply for obtaining records? How will late or missing records be dealt with? Will amendments to the data be made overtime (will revisions to series be made?)
- Do the data items collected adequately fit within the reference period? Has the reference period been consistently applied to the collection of all data?
- How often will the data be extracted / transferred to the receiving agency for use?

- *Accuracy*
  - Have the data been adjusted in any way? e.g. imputed, substituted, edited. If so, what adjustments were made?
  - What is the collection size?
  - What are the rates of non-response or under reporting?
  - Have any parts of the population been unaccounted for?
  - What has been done to manage under counts or over counts, if any were present?
  - Are there sensitive or biased questions or topics collected? e.g. age or income.
  - Are there hierarchical records? Are all levels of the hierarchy required or do the records require some transformation into a flat record? Do the records require some grouping or consolidation of reporting units into a statistical unit?
  - Can the receiving agency request the data custodian to do any follow up work in regards to missing data?
- *Coherence*
  - Are the data consistent over time and across organisations (i.e. can data from different jurisdictions be compared or are the data incomparable)?
  - Are rates or percentages calculated using the same data source? If not, then further information needs to be provided on how these figures have been created (i.e. what differences affect the comparability and what impact do the difference have?).
  - What standards and classifications of data items are used?
  - Is a time series of the data available?
  - Have there been changes to the underlying data collection?
  - Are there any real world events (e.g. new laws, tourist events or disease outbreaks) that could impact on the data?
- *Interpretability*
  - Are there any contextual issues that need to be raised?
  - Is other information available to help users understand the data source?
  - Are there ambiguous or technical terms that may require further explanation?
- *Accessibility*
  - Can data that are not released be requested? If so, what is the correct process for the request?
  - What format are the data available in? e.g. spreadsheet, column specifications etc.
  - Are there any privacy or confidentiality issues to be considered?

ASSESSING A DATA SOURCE *continued*

- What security will be in place for the transfer of the information from the data custodian to the receiving agency?

For more examples of questions for assessing administrative data sources please see " http://www.nss.gov.au/dataquality/PDFs/DQO_Admin.pdf" on the National Statistical Service website.

In conjunction with the list of questions for assessing administrative data sources at the acquisition stage, it would be useful to also compile a list of quality measures (or indicators) that will help in an assessment of the data. Once again using the *ABS Data Quality Framework* as a guide to ensure all aspects of quality are measured. These measures can be qualitative (e.g. Yes / No) or quantitative (e.g. 97%). This assessment could be repeated each time the data custodian provides the administrative data to the receiving agency in order to provide the most up to date quality information about the statistics. This information may then be included in the report that accompanies the statistical outputs derived from the administrative data.

Some examples of quality measures for assessing a data source could include, but are not limited to:

- *Institutional Environment*
  - Changes in the data custodian's underlying legislation;
- *Relevance*
  - Frequency distribution for key variables;
  - Outstanding returns;
  - Coverage of population;
- *Timeliness*
  - Date of receipt of records;
- *Accuracy*
  - Number of missing values for key variables;
- *Coherence*
  - Population counts (which can be monitored overtime by themselves and/or against predicted likely counts);
  - Business births since last period;
  - Business deaths since last period;
- *Interpretability*
  - Changes to units;
  - Changes to definitions; and
- *Accessibility*
  - Number of new staff directly involved in the managing of the administrative data at the data custodian agency.

CHAPTER 4 MANAGEMENT OF RELATIONSHIPS WITH DATA CUSTODIANS

MANAGEMENT OF RELATIONSHIPS WITH DATA CUSTODIANS

The most important strategy for managing a range of risks with the use of administrative data is the maintenance of ongoing relationships with the data custodians of the source data. These relationships should be established to varying degrees at all levels from the Chief Executive down to operational staff.

It is always a sound principle to maintain good relationships with the data custodians of information to be used for statistical purposes. Unlike the direct collection of information the use of administrative data as a source introduces a third party into the supply chain. Sometimes even a fourth party if the data custodians of the source data contract out the management of its day to day operation. These additional parties have other obligations and interests and for them the production of statistics from their source data will often be of secondary or lower importance.

Administrative data systems are usually set up by, and operate under, other legislation with specific objectives to be delivered by the data custodian. In addition, more general legislation relating to privacy also applies. The best way forward is for the various agencies involved to work together to achieve their respective objectives in accordance with their respective legislation.

In any successful relationship, the foremost principle is to understand the needs of the data custodians (and managers) of the administration data. However, it is important to make sure the data custodians understand the value of the statistical information which can be produced from their data and the critical requirements of the data for producing statistics of the required quality. This is especially important if the contact person changes as relationships will need to be re-established. Also, priorities and perspectives will need to be re-discovered or at least checked to see if they are consistent with the previous contact. Due to the large number of data custodians that have contracted out their information technology work, developing a good relationship with these contractors is important as they may not be aware of the quality of the data provided. Hence, helping them to understand the statistical aspects of the data and how it is used as well as what is done with it, will be useful to these contractors.

MANAGEMENT OF RELATIONSHIPS WITH DATA CUSTODIANS *continued*

The following is a list which has been compiled from the experiences of the ABS in establishing and maintaining sound relationships with data custodians:

- Establish governance and rules;
- Provide advice and help; and
- Maintain communication.

Establish governance and rules

- Use some form of written agreement between the parties. These agreements have come in a variety of forms over the years but essentially set out objectives and obligations for both sides. Memorandums of Understanding (MOUs), Data Service Agreements, and Service Level Agreements (SLAs) are examples of agreements used. An SLA or MOU covering what is collected and in what form etc. is good practice for both parties involved. The SLA or MOU should specify what, when and how information is to be provided, the confidentiality requirements of both the data custodian and the receiving agency, the security requirements around the handling of the data, etc. It should have quality requirements specified and obligations on the part of parties should these requirements not be met.
- Establish some form of governance arrangements involving senior managers, at least for critical stages of projects such as implementation of a new or revised system. Consider involving wider relationships by involving other stakeholders with an interest in ensuring successful utilisation of the source data. The arrangements should provide for the management of incidents which can impact on supply and quality.

Provide advice and help

- If a new administrative source is being established which will be of value statistically then the ideal situation is for the receiving agency to be involved with the project right from the start so statistical aspects can be built into the system being implemented. This however is rarely the case as most systems have been around a long time. The next best option is for the receiving agency to be involved when the administrative systems are being redesigned to meet changed objectives or to operate under new information technology. This ensures that statistical requirements can be considered and catered for.
- Understand what expertise from within the receiving agency can bring to bear on a project or to support ongoing operation. This expertise includes statistical coding software, information management skills, statistical standards for collecting and using certain information (e.g. demographic information about a respondent, geographic classification).
- Establish a champion for statistical use within the data custodian agency. This could be a senior executive or a key user of the statistics (e.g. policy evaluator associated with the administrative system). Sometimes the information is so critical that agreements need to be reached between the respective chief executives.
- Providing feedback to the data custodian at the end of each cycle is a useful process as it offers the opportunity to inform the data custodian about how their own processes can be improved. This will encourage the data custodian to use the information given to improve upon their own internal process and hence improve the data provided for future cycles. Regular reviews of on-going collections are important.

MANAGEMENT OF RELATIONSHIPS WITH DATA CUSTODIANS *continued*

Maintain communication

- Ensure the data custodians understand the value and importance of the key statistics produced from their source. This includes ensuring they understand the key quality requirements such as accuracy and timeliness of supply, and of course the impact of disruption of supply.
- Maintain continuous contact with the data custodian through email, telephone, in person etc. This contact should not be limited to when the receiving agency requires something from the data custodian.
- Internal relations also need to be maintained to ensure that there are not multiple people from the receiving agency all vying for the attention of the data custodian.
- Map not only the data custodian of the data but also other users of the data who may be able to influence the data custodian. This is beneficial when changes might need to be made to the existing administrative dataset e.g. the addition of extra questions, wording changes etc. The receiving agency can seek assistance from the influential users to assist with getting potential changes implemented to the administrative collections where there are mutual benefits for the changes.

CHAPTER 5 MANAGING RISKS

MANAGING RISKS

There are numerous issues to be aware of with the acquisition of administrative data, some of which have been covered. Along with determining the data requirements in order to assess administrative data sources prior to acquiring them it is important to think about the monitoring of the administrative data that is needed. This is to ensure its quality from the collection of the information by the data custodian, the hand over of the information and the subsequent processing of the data by the receiving agency.

In each of these scenarios it may be useful to have a monitoring device in place which reports timely information in order to assess whether the quality of the data and process are meeting the quality requirements.

The ABS uses a tool called Quality Gates to help manage statistical processes. Quality gates consist of a set of acceptance criteria imposed at predetermined points in a production process whereby each of the components (Placement, Quality Measures, Roles, Tolerance, Actions and Evaluation) play an important part in determining the fitness for purpose of the process. More information on the detail, use and implementation of quality gates can be found in the information paper "*Quality Management of Statistical Processes Using Quality Gates, Dec 2010*" (cat. no. 1540.0)".

Further information on managing the statistical processes of administrative data in a process cycle will be covered in detail in a future publication.

Administrative data is a useful resource for substituting and replacing direct collection of data. However, there are various issues that need to be considered and managed prior to its use in a receiving agency. Provided the receiving agency is aware of the issues and manages them appropriately then administrative data is a useful source of information which can save time, money and reduce respondent burden.

BIBLIOGRAPHY

REFERENCES

ABS (Australian Bureau of Statistics) 2009a, *ABS Data Quality Framework, May 2009* , cat. no. 1520.0, ABS, Canberra. Website address: http://www.abs.gov.au/ausstats/abs@.nsf/mf/1520.0

ABS 2009b, *Quality Manual, Statistics Sourced from Administrative Data*, ABS, Canberra

ABS 2009c, Research being undertaken in macroediting, *Methodological News, Mar 2009*, cat. no. 1504.0, Canberra. Website address: http://www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/1504.0Main+Features10Mar+2009

ABS 2010, *Quality Management of Statistical Processes Using Quality Gates, Dec 2010*, cat. no. 1540.0, ABS, Canberra. Website address: http://www.abs.gov.au/ausstats/abs@.nsf/mf/1540.0

Michaud, S., Dolson, D., Adams, D., and Renaud, M. (1995). Combining administrative and survey data to reduce respondent burden in longitudinal surveys. *Proceedings of the Section on Survey Research Methods*, American Statistical Association. p.11–20.

NSS (National Statistical Services) 2010, *Define a Data Need*, NSS, Canberra. Website address: http://www.nss.gov.au/dataquality/PDFs/DQO_Needs.pdf

NSS (National Statistical Services) 2010, *Data Quality Online Questions : Administrative Data*, NSS, Canberra. Website address: http://www.nss.gov.au/dataquality/PDFs/DQO_Admin.pdf

ONS (Office of National Statistics) 2005, *Atkinson Review: Final Report, Measurement of Government Output and Productivity for the National Accounts*, 2005, Crown Copyright. Website address: http://www.statistics.gov.uk/about/data/methodology/specific/PublicSector/Atkinson/downloads/Atkinson_Report_Full.pdf

Public Works and Government Services (2007), *Best Practices in Public Opinion Research: Improving Respondent Cooperation for Telephone Surveys*, Public Works and Government Services Canada, Ottawa. Website Address: http://www.tpsgc-pwgsc.gc.ca/rop-por/rapports-reports/telephone/etape-stage-01-eng.html

SDMX (Statistical Data and Metadata eXchange) 2009, *Annex 4: Metadata Common Vocabulary*, 2009, SDMX. Website address: http://sdmx.org/wp-content/uploads/2009/01/04_sdmx_cog_annex_4_mcv_2009.pdf

FOR MORE INFORMATION . . .

INTERNET **www.abs.gov.au** the ABS website is the best place for data from our publications and information about the ABS.

INFORMATION AND REFERRAL SERVICE

Our consultants can help you access the full range of information published by the ABS that is available free of charge from our website. Information tailored to your needs can also be requested as a 'user pays' service. Specialists are on hand to help you with analytical or methodological advice.

PHONE 1300 135 070

EMAIL client.services@abs.gov.au

FAX 1300 135 211

POST Client Services, ABS, GPO Box 796, Sydney NSW 2001

FREE ACCESS TO STATISTICS

All statistics on the ABS website can be downloaded free of charge.

WEB ADDRESS www.abs.gov.au