# NSS Handbook

# Australia

Also available at: http://www.nss.gov.au/nss/home.nsf

**Chapter 1 - Introduction**

## 1.1 PURPOSE AND STRUCTURE OF THE HANDBOOK

The purpose of this handbook is to outline the issues that need to be addressed when:

- conducting a statistical collection,
- extracting data from administrative systems,
- managing statistical data, or,
- turning administrative or survey data into statistics and their subsequent analysis.

This handbook is a reference guide, providing a broad overview of the statistical process to the inexperienced collection manager, designer or user of statistical products. Each element of the handbook introduces a different aspect of the statistical cycle, including, where appropriate, references to more advanced sources of information.

Chapters 2 to 8 introduce each phase of the statistical cycle. These are:
2. Initiating the Statistical Activity
3. Designing the Statistical Cycle
4. Data Collection and Data Extraction
5. Data Processing
6. Statistical Analysis and Interpretation
7. Dissemination of Statistics
8. Evaluation of the Statistical Activity

See Appendix 1 - Phases of the Statistical Cycle for a detailed outline of the phases of the statistical cycle.

The Appendices cover issues and considerations that transgress all phases of the statistical cycle. These are:

1. Phases of the Statistical Cycle
2. Data Management
3. Standards and Classifications
4. Confidentiality and Privacy
5. Information Development Plans
6. Statistical Skills Development
7. Project Management
8. Quality Declaration

The information presented in the NSS Handbook is directly relatable to the NSS Key Principles. The Principles articulate 'best practice' guidelines to promote the achievement of high standards in the collection, compilation and dissemination of statistics. The objective of the NSS Key Principles is to assist government organisations, at Australian, State/Territory and local levels, in their statistical activities.

## 1.2 STATISTICAL SURVEY COLLECTIONS AND ADMINISTRATIVE DATA SYSTEMS

This Handbook is intended to apply to data from both statistical survey collections and administrative systems.

Statistical surveys collect information in order to answer a question or to make a decision by placing a value onto an indicator or measure. Survey collections measure one or more characteristics of a population. These characteristics may be measured by surveying all members of the population, or a sample of the population.

An administrative system holds data primarily for the purpose of carrying out various programs, for example, school enrolment applications, hospital separation records, birth, death and marriage certificates and income tax collection. The range of information held is mainly determined by what information is required to manage the particular program.

Although the purpose of an administrative data system is for direct use in the administration of a program, this does not preclude the use of such data for statistical purposes. In fact, administrative data is a powerful source of basic data. These data can be summarised to provide statistical information for the population related to the program. The data from administrative collections can however be difficult to interpret as statistical information. This problem can be addressed either by improving the extraction of statistics from the existing administrative data collections and/or using the administrative data collection vehicle to collect additional statistical data for the specific purpose of providing statistical information.

The underlying principles governing good practice in statistical collections are generally equally applicable to the extraction of data from an administrative system. The chapters in this Handbook cover the statistical cycle from planning statistical activities to disseminating and evaluating statistical information, while also describing other requirements for creating useable data, including protecting confidentiality, managing data effectively and using standards.

Initiating the activity (Chapter 2) and designing the activity (Chapter 3) are necessary steps in both setting up an administrative system and undertaking a statistical collection. The data collection and extraction (Chapter 4) and processing (Chapter 5) guidelines will apply to administrative records only if they exist in hard copy and need to be coded and captured. Editing (Chapter 5) and dissemination (Chapter 7) guidelines apply to all cases of administrative data where a file of individual administrative records is obtained or created for subsequent statistical purposes.

Regularly reviewing and monitoring administrative and statistical systems ensures that a high level of data quality is maintained (Chapter 8). The production of high quality data also involves establishing good data management practices (Appendix 2) and applying standards wherever possible when designing administrative and survey forms and producing statistics (Appendix 3).

Analysis and interpretation of statistical data and data extracted from administrative systems assists the data to be more widely accessible and useful. Data analysis and interpretation allows raw data to be reported in a way that assists to highlight key population characteristics of survey results. For administrative data, this can involve examining whether a program is effective or examining characteristics of the population to which the program relates (Chapter 6).

Evaluation (Chapter 8) is a key component of data collections, and of data extraction systems for administrative data, to allow for improvements in operations and systems.

The techniques for confidentialising statistical information produced from administrative data (Appendix 4) will sometimes be different to the techniques for confidentialising tabular data since the data from administrative records are often available at a very detailed level. The legislative obligations governing privacy will also have implications for agencies developing administrative systems and undertaking data collections.

# Chapter 2 - Initiating the Statistical Activity

## OVERVIEW

Initiating a statistical activity covers developing/redeveloping a new or existing survey or census, developing a new administrative data system, or developing a process to extract statistical information from an existing administrative data system. This chapter identifies the main elements to be considered in initiating the statistical activity. It covers issues relating to developing objectives, budget and resources, timeliness, existing data sources, sensitivity, project management, analysis and reporting and research.

## 2.1 OBJECTIVES

### 2.1.1 Developing the Objectives with Stakeholders

Creating a comprehensive list of the key users and owners of the data (stakeholders) will assist to target consultation to develop a full set of objectives for the statistical activity. A common problem with statistical activities is a narrow focus when developing the objectives which limits the usefulness of the resulting data. Often, wider consultation at the initial planning stages can enhance the useability of data with little additional cost.

Potential stakeholders include both internal and external users. Internal stakeholders typically include people from a wide range of areas within the agency. For example, it is often easy to overlook some important corporate objectives if the focus is kept within a particular section, branch or division. An agency involved in making policy decisions would be required to meet the needs of its Minister. External users may play an important role in informing and influencing future policy directions through their data requirements.

Similarly, it can be valuable for an agency setting up an administrative system to consider all the statistical uses of that system within and outside their own agency. By enlisting external stakeholders to develop the objectives for an administrative system, planning will include decisions on how to make the information available, in some manner, to external users.

### 2.1.2 Defining the Objectives

Poorly defined objectives will often lead to statistical activities that deliver data which do not meet the desired uses of the stakeholders and can lead to inefficiencies in the design of the activity. Well defined objectives include a description of the concept(s) to be measured, the population of interest and the level of accuracy required. These are largely determined by the intended use of the statistical data and will impact greatly on the cost associated with collecting or extracting the data.

For example, if you were collecting information on income it could be collected in a number of different manners:

- as a broad, cross classificatory variable expressed in a small number of broad ranges; or
- as a detailed dollar figure collecting only wage and salary income; or
- as detailed variables indicating dollar figures for all forms of income.

These different forms of income are collected to meet different intended uses of the data and have considerably different costs associated with their collection.

### 2.1.3 Prioritising the Objectives

The planning phase deals largely with balancing the requirements of the user/users with the constraints of budget, resources and time available. This balancing process is made easier when the objectives are prioritised since it is usually the case that not all objectives will be able to be met by a single collection or process.

As more detailed development is undertaken it will become clear whether the number of objectives for the statistical activity need to be reduced or can be expanded.

### 2.1.4 Information Development Plans

Developing the objectives of a statistical activity can be greatly enhanced through the use of Information Development Plans. The aim of Information Development Plans is to identify and meet users' needs for statistics within a particular field through a strategic process. They identify and develop strategies to address statistical gaps and deficiencies in the frameworks, classifications and standards in a particular field of statistics. Information Development Plans are usually developed jointly by stakeholders within the field of interest and provide guidance for all statistical activities within the field.

Where Information Development Plans exist within the field of interest any proposed statistical activity should be guided by the plan. See Appendix 5 - Information Development Plans for further details.

## 2.2 BUDGET AND RESOURCES

### 2.2.1 Budget

The main constraint to a statistical activity is usually the available funding. At the early stages of planning it may be difficult to estimate costs for the statistical activity with any degree of accuracy. Indeed, the costs are largely dependent upon choices such as whether a new survey needs to be developed and conducted or if the data can be extracted from an administrative system or obtained from an existing survey. Other decisions, such as the type of survey to develop, are also highly dependent upon the budget available. It is often necessary to propose several options for the statistical activity, each requiring varying budgets and having different restrictions.

When estimating costs it is very important to consider all resources which will be required. These usually fall into three categories, outlined below. However, if the agency engages consultants, these costs will be subsumed in the overall consultancy cost.

- Salaries - include both office staff and any temporary or field staff that may be required. Also consider whether you need to include full staff costs (i.e. include superannuation, information technology and accommodation costs) or just direct costs (i.e. only the wage component).

- IT costs - include computing costs beyond the normal staff information technology costs, such as software costs, information technology infrastructure development and running costs, and data storage costs.

- Administrative costs - including consultancy costs, training, printing of survey forms, postage, travel expenses, publication and dissemination costs and the purchase of lists (for the survey frame).

### 2.2.2 Resources

As well as budget constraints, the availability of resources also needs to be considered. A statistical activity may require specialised resources, whether they be staff skilled in survey design or computer programming, or the availability of particular information technology resources. Where these resources are in scarce supply and/or heavy demand it may not be just a case of having the money to pay for them.

The availability of some resources may affect the timing of the activity or may mean that the activity needs to be developed to use other available resources. Resources may become a particular concern if the statistical activity is required to be conducted within a very short time frame.

### 2.2.3 Training

Staff need to be provided with the necessary training and tools to acquire appropriate skills for the effective management of statistical activities. Agencies should assess staff skill needs in relation to their statistical activities, so that any deficiencies can determined and strategies put in place to address them. Any skill deficiencies can be addressed by recruiting appropriate staff, providing training and development opportunities, obtaining assistance from consultants or obtaining expert advice from other areas in the organisation with specialist skills. More information can be found in Appendix 6 - Statistical Skills.

## 2.3 TIMELINESS

Users usually want data as quickly as possible i.e they need to be timely. A process which is only able to deliver data several years after they have been collected is not likely to meet many of the needs of the users of the data.

The time frames for the statistical activity should be consistent with the objectives, for example, is the data required for policy development or to monitor the effects of changes in policy. Timing the availability of statistical data to coincide with such events may have a major impact on the scope of the activity if only limited time is available.

The length of time required to undertake a statistical activity can often be shortened by making more funds and resources available. However, as funds and resources are usually limited, the objectives of the activity may need to be reduced to facilitate a shorter timetable.

## 2.4 EXISTING DATA SOURCES

Before embarking upon developing a new statistical collection or when reviewing the need for an existing collection an assessment should be made of the usefulness of any existing data. Much of the information which organisations require may be obtained from sharing data from the administrative systems or statistical collections of other organisations or the information may already be available from within the organisation itself.

Many administrative systems collect a great deal of information about the population within the program the system was developed to administer. Administrative data often covers the entire population within the program and the information is usually up to date. Such a rich source of data is very expensive to duplicate in a survey or census. Using existing data can be a timely and cost-efficient solution provided the existing data meet the requirements of the research.

When considering whether to use existing data it is important to understand how and why the data has been collected. Administrative systems collect information which is necessary to administer particular programs or processes. If statistical applications of the data have not been considered in developing these systems, then the concepts and standards used to collect and process the information may result in the data not meeting the needs of the proposed statistical activity. It is also necessary to understand the legislative and confidentiality issues related to using administrative data for a purpose for which it was not initially collected. See Appendix 4 - Confidentiality and Privacy for further information regarding legislative and confidentiality issues.

In addition to the Australian Bureau of Statistics, Commonwealth and State/Territory Government agencies, local councils, universities and other research agencies are sources of published statistics. Government agencies and professional or business associations can be good sources of unpublished information. Sometimes a search of existing data will not precisely satisfy the information need, but it might give data related to it or part of the information required. This information might be adequate for research purposes.

### 2.4.1 Data Quality Framework

A specific set of criteria or framework should be used to assess the suitability of existing data sources. The Australian Bureau of Statistics uses a data quality framework to assist in evaluating whether data are fir for purpose. The criteria within the framework are:

- Relevance

The relevance of statistical information reflects the degree to which it meets the real needs of clients. It is usually described in terms of key user needs, key concepts and classifications used and the scope of the collection (including the reference period).

- Accuracy

The accuracy of statistical information is the degree to which the information correctly describes the phenomena it was designed to measure.
It is usually characterised in terms of error in statistical estimates and is traditionally decomposed into bias (systematic error) and variance (random error) components.

- Timeliness

The timeliness of statistical information refers to the delay between the reference point (or the end of the reference period) to which the information pertains, and the date on which the information becomes available.

- Accessibility

The accessibility of statistical information refers to the ease with which it can be referenced by users. This includes the ease with which the existence of information can be ascertained, as well as the suitability of the form or medium through which the information can be accessed.

- Interpretability

The interpretability of statistical information reflects the availability of the supplementary information and metadata necessary to interpret and utilise it appropriately. In addition, it includes the appropriate presentation of data such that it aids in the correct interpretation of the data.

- Coherence

The coherence of statistical information reflects the degree to which it can be successfully brought together with other statistical information within a broad analytic framework and over time. The use of standard concepts, classifications and target populations promotes coherence, as does the use of common methodology across surveys.

For more information see Appendix 7 - Quality Declaration and Assessment.

## 2.5 SENSITIVITY

The collection and use of sensitive data, or data considered sensitive by respondents, can provoke adverse respondent reactions and jeopardise response rates to surveys. The time taken to collect sensitive data is generally greater than non-sensitive data which can mean that the time available to collect other data is reduced.

Development time and costs for this type of survey are likely to be greater as more detailed testing of the effects of collecting the sensitive data items needs to be undertaken.

Being well informed of the potential sensitivity of a collection to the respondent will assist to manage the provider relationship. Assurances of confidentiality and good collection design act to improve response rates to sensitive information.

### 2.5.1 Legislation

Restrictions associated with confidentiality provisions and Privacy legislation must be taken into consideration in determining whether data can be utilised for statistical purposes, or purposes other than that for which it was originally collected. Administrative systems clients are often informed that the unit record data will only be available to a limited number of agencies.

Where a file containing information enabling the individuals concerned to be identified is released to other individuals, other agencies or the general public, active management of the inherent risk that the privacy and security of individuals' information may be compromised is required. The web site of the Office of the Federal Privacy Commissioner (http//www.privacy.gov.au) outlines the limits on the disclosure of personal information under the Information Privacy Principles contained in the 'Privacy Act 1988'.

Information Privacy Principle 2 requires collection forms to make it clear that the personal information may be passed on to certain specified agencies for a range of purposes (e.g. evaluation and monitoring of the program). Personal information is defined as "information or an opinion.......about an individual whose identity is apparent, or can reasonably be ascertained, from the information or opinion".

The non-routine provision of personal information to other departments, organisations, or individuals can only occur if it meets the requirement of Information Privacy Principle 11 which deals with limits on disclosure of personal information. This principle severely limits the scope for extending access to these data beyond the list of nominated departments or contractors engaged directly by them under specific conditions. While the data may be de-identified by the removal of names, addresses and administrative reference numbers, they may still contain detailed data, including regional data that could in some cases be used to identify the individuals who provided them. They would therefore fall within the definition of personal information and the constraints of Information Privacy Principle 11.

A useful source of guidance is to review existing legislation and principles that relate to your own agency. More information in relation to relevant Privacy Principles can be found in Appendix 4 - Confidentiality and Privacy. Departmental principles and guidelines also inform the application of these principles in specific agencies.

## 2.6 PROJECT MANAGEMENT

Project management in the initial planning stage should be used to ensure collections and administrative data systems are well managed, and a balance is achieved with respect to the objectives, budget and resources, and timeliness. Usually the objectives and the budget are the two main parameters which need to be balanced. Often at least one of these parameters will be largely set and the decision will then necessitate aligning the other parameter with this. For example, if there is a limited budget then the objectives need to be brought into line with this available budget.

Project management for statistics also involves the capacity to effectively use committees and working groups, service level agreements (setting and agreeing to timetables and quality goals), and cater for a survey review and evaluation process. More information about Project Management can be found in Appendix 8 - Project Management.

## 2.7 ANALYSIS AND REPORTS

Decisions should be made on the analysis to be conducted and reports to be produced when initiating the statistical activity. Expert assistance may need to be sought about what particular analytical tools are appropriate. Reports covering survey results may range in style from a comprehensive report to a condensed summary report supported only by significant graphs or tables. In all cases, however, it is vital that details of the survey methodology (sample selection, method of information collection, quality control etc) are documented because the particular methodology used can affect the analysis and interpretation of results. Analysis and reports are discussed further in Appendix 6 - Statistical Skills.

## 2.8 RESEARCH

Research should be undertaken at an early stage to become familiar with the issues related to the planned statistical activity. It should cover issues associated with the actual subject matter or topic about which statistical data are to be collected and issues relating to the statistical process.

Research about the subject matter should aim to identify particular issues related to the planned use of the statistical data. These may include:

- understanding the context and issues surrounding the objectives and their implications for statistical measurement
- relevant research in the area
- talking to experts in the field of interest
- existing Information Development Plans (see Appendix 5 - Information Development Plans)
- accepted standards and concepts
- areas where there is no available data or available data is deficient or out of date
- changes which have occurred within the area of interest which are likely to impact on the analysis of data (e.g. legislative changes resulting in definition or requirement changes, introduction of a new classification, and/or changes to procedures).

Research into the statistical process aims to identify issues which will affect the manner of collection, processing and analysis of the data. These may include:

- what data are currently available, from previous surveys or from existing administrative systems?
- identify other sources of data to which it may be useful to link. This is important to consider at an early stage in terms of data items, classifications and standards.
- the need for comparability with previous data
- reports from previous statistical activities on the subject matter.

**Chapter 3 - Designing the Statistical Activity**

## OVERVIEW

This chapter outlines the major issues to be considered when designing a survey or an administrative system. It examines topics relating to survey design, data collection, sample methodology, questionnaire development and testing, and respondent load.

## 3.1 SURVEY DESIGN

The overall design of the statistical activity, whether it be a survey or an extraction from administrative data, should consider all aspects from collection of the information through to dissemination and use of the statistical data. The various parts of the statistical activity are highly interrelated and the overall design should provide a framework to ensure all stages link together efficiently and satisfy the overall objectives of the activity.

### 3.1.1 Data Requirements

Detailed data requirements should be specified early in the process, as changing the design of a collection to incorporate new data items or revised data items can be costly. Well defined data requirements will ensure sufficient time to properly test new items and correct potential data quality issues.

Data items should reflect the overall objectives of the activity, rather than the actual questions to be asked. Data requirements should be expressed in terms of what concept is being measured and what data will be the resultant output. Specifications about the data should not be too general, as this may lead to user needs not being met.

Data specifications, together with resource constraints, will largely determine whether existing data sources, e.g. administrative data or an existing collection, can be used or if a new collection is required. Further, the data specifications will largely drive the choices which are made in relation to collection methodology, sampling methodology, processing and analysis. Resource constraints alone might determine that existing data sources have to be used.

Some decisions on the design of a survey may make it impossible to provide certain data. For example, if a sampling methodology is developed to provide statistical estimates at a State level, the resulting output may not allow meaningful analysis of issues at lower geographic levels.

### 3.1.2 Units of Analysis

Units of analysis need to be defined during development of the collection to determine which data items in relation to the units about which the agency intends to collect information. They can be physical units such as persons, families, households, business units or events such as births, hospital separations, or transactions such as sales.

### 3.1.3 Population

The population is the aggregate or collection of units about which the survey will be conducted. Units can refer to people, households, schools, hospitals, businesses. There are two different populations with which a survey is concerned:

- The target population or scope of the survey is the group about which inferences would like to be made from the survey data. The target population should be defined in terms of content (eg all persons), units (eg in households), extent (eg in Australia) and time frame (eg 31 December 2003). The target population may not be easily accessible due to logistical or financial constraints. For example, in the Census of Population and Housing, it is difficult to obtain data on the homeless or people who are in transit from one place to another during Census day.

- The coverage or survey population is the actual set of units about which information can be obtained or inferred. The survey population should accurately represent the target population so that the survey results can be used for their intended objectives.

### 3.1.4 Data Items

The outputs should be identified during the planning phase to ensure that the content of the collection is appropriate. Output tables should be created relatively early in the process to ensure that the collection methodology will meet survey or research objectives.

Data items determined by the objectives for the collection form the basis for question wording. Data items need to be carefully converted into questions and each data item should be thoroughly tested. The data items, along with metadata should be linked with the data management strategy. See Appendix 2 - Data Management for further information.

### 3.1.5 Standards and Classifications

In the planning stage of the survey, the data items to be collected should be clearly defined according to the relevant standards and classifications. The use of standard frameworks for presenting, collecting and using data facilitates greater application of the data. For example, in a survey of businesses, respondents may need to be classified according to their industrial activity. If this is done using a standard format, the resulting statistics can be compared with existing published statistics or with a later repeat of the survey. Likewise, in a household survey, if income is collected using the standard definition, it can be used on a comparable basis. Similar conditions can be applied to classifications for geographic region.

By making use of standard concepts and data items it may also be possible to integrate data from different organisations. A number of standards have been developed in keeping with this theme. This integration improves the comparability and relevance of the data and reduces duplication of data collection.

Standard frameworks may use a combination of data items to produce the required output item. These outputs are referred to as derived items. For example, the unit of analysis for a survey of the labour force would be persons, and a data item collected would be hours worked. The labour force framework classifies, at a given moment in time, the population aged 15 and over for measuring the population into three categories: employed, unemployed and not in the labour force. The concept of labour force is a derived item and uses the employed and unemployed categories together.

Appendix 3 - Standards and Classifications provides more information about this topic.

## 3.2 DATA COLLECTION METHODS

The success of the survey will depend to a large extent on the suitability of the collection method chosen. The balance between the objectives (or data requirements) and the resource constraints may limit choices for survey methodology. Other factors to consider are:

- Complexity of topic and nature of questions; issues include requirements for respondents to access records, the need to provide detailed explanations to respondents, complex sequencing required for varying population groups within the survey, and the inclusion of questions on sensitive topics.

- Response rates; although it is very rare to achieve a 100% response rate for any survey, the choice of collection method can significantly influence the response rate obtained. For example, personal interviews usually achieve a better response rate than mail surveys or telephone surveys.

- Respondent preference; a collection method which fits in with the life style or working style of the respondent should be considered. Some respondents have a strong preference for completing their forms electronically, and the Electronic Transaction Act (July 2001) allows respondents to insist on completing surveys using this method.

- Sampling frame and target population; certain collection methods will not be suitable for certain target populations. For example, a mail or phone survey would not be suitable for a survey of homeless youth.

The most common collection methodologies are described below. It should be noted that it is rare that one methodology will be the best solution to all issues. The methodology which provides the best overall solution should be selected or alternatively a mix of different methodologies. For example, a face to face interview could be supplemented with a self enumeration form to enable respondents to complete sensitive details they may not be comfortable telling an interviewer.

### 3.2.1 Face to Face Interviews

Face to face interviews are used mainly for household surveys. Personal interviews involve a trained interviewer going to the potential respondent, asking the questions and recording the responses.

**Advantages**

- enables the interviewer to explain the purpose of the collection and may improve the response rate

- enables the interviewer to provide explanations as required and use interview aids such as show/prompt cards
- allows longer interviews than telephone
- allows surveys to use complex sequencing.

**Disadvantages**

- expensive, training and travel are substantial costs for complex face to face interviews
- data can be subject to interviewer bias (caused by interviewer's appearance, attitude, phrasing of questions and explanations)
- respondents may be reluctant to disclose sensitive or private information to interviewers

- interviewers with appropriate language skills are required for some respondents who are not proficient in English.

### 3.2.2 Telephone Interviews

Telephone surveys are used for both household and business surveys. Using this method a questionnaire is administered over the phone and responses are recorded manually or directly into a computer system by the interviewer. Telephone data collection methods are also widely used for follow-up and post-enumeration work for other collection methods.

**Advantages**

- reduced costs, compared to personal interviews as it is possible to cover all Australia from one call centre
- the interviewer can explain the importance of the survey and establish rapport with respondents, improving response rates
- enables the interviewer to provide explanations as required, and conduct complex sequences of questions
- produces timely results
- call-backs for people 'not-answering' and follow-ups for additional information are relatively quick and inexpensive.

**Disadvantages**

- limited in number and complexity of questions that can be asked
- because of ease with which the respondent can terminate the interview, non-response and partial non-response can be higher than face to face interviews
- bias is introduced because people with no phones or who are rarely available are excluded, specialist methodologies have to be used to produce a reliable sampling frame because of unlisted numbers and changes in addresses
- problems with convincing respondents of the authority behind survey and confidentiality of results.

### 3.2.3 Self-Enumeration Methods

**Postal/mail survey**

Copies of questionnaires are mailed out to respondents with a reply-paid envelope so that the respondent can mail back the completed form. Follow up procedures are also usually conducted by mail.

**Advantages**

- cheaper than personal interviews
- respondent is able to complete questionnaire in their own time
- respondent can check records
- detailed instructions and explanations can be included for improved data quality
- allows access to 'difficult-to-contact' respondents.

**Disadvantages**

- generally has a lower response rate because it is easy for respondents to discard questionnaire
- time lag between when questionnaire is mailed out and the time it is returned
- limit on complexity of questionnaire compared to face to face interviews
- respondent may misinterpret instructions
- not appropriate for potential respondents with limited ability to read or write English unless questionnaires are provided in appropriate language.

**Drop off Mailback and Drop off Pick-up**

The questionnaire is delivered to respondents by an interviewer who explains the aims of the survey and how to fill out the questionnaire. The questionnaire is left with the respondent to be completed and either mailed back by the respondent or picked up by the interviewer at a later date.

**Advantages**

- generally provide higher response rates than postal surveys
- usually less expensive than personal interviews.

**Disadvantages**

- costs involved in using interviewers makes this technique more expensive than postal surveys
- respondent has to be available when interviewer visits.

## 3.3 DATA COLLECTION TOOLS

The different collection methods can be either paper based or electronic based or a combination of the two. An electronic based survey can involve either computer assisted interviewing or electronic forms.

### 3.3.1 Computer Assisted Interviewing

Interviewer uses a computer to enumerate the survey rather than a paper form. Can be used for both face to face (Computer Assisted Personal Interviewing or CAPI) and telephone interviewing (Computer Assisted Telephone Interviewing or CATI).

**Advantages**

- sequencing of questions is controlled by the computer allowing complex sequencing
- allows some editing and querying of responses to be carried out at time of interview, improving data quality
- data entry and some coding and processing are part of the interview process which improves the efficiency of the survey
- 'call scheduling' for telephone interviewing allows calls to be rescheduled if the phone is engaged or respondents are not available
- allows interviewing staff to have their performance monitored.

**Disadvantages**

- increased cost in set-up and maintenance of computer equipment and training of interviewers.

The following paper, SCH web site, Review of CATI Procedures in Overseas Statistical Agencies, provides some information on overseas experiences with CATI systems.

### 3.3.2 Electronic Form/ Computer Assisted Self Interview/ Internet Surveys

An electronic form (eform) is an electronic version of the questionnaire that can be sent to the respondent's computer via e-mail or accessed from the World Wide Web.

**Advantages**

- data entered onto an eform can be edited and responses queried at point of entry, improving data quality
- use of electronic returns produces faster response than other self-enumeration methods
- reduction in labour costs as no interviewers are required
- questions can be sequenced so only questions relevant to the respondents are visible
- scope for interactive questionnaires and complex skip patterns.

**Disadvantages**

- increased cost for development of forms
- maintenance of related systems and security
- relies on widespread internet access in the community
- coverage errors as only those interested in a subject may access required links
- absence of well-established design standards
- requires respondents to have comparable computer software and help-desk type staff may be necessary to support use of the form.

## 3.4 SAMPLE METHODOLOGY

Once the survey design and data collection method have been determined, the sample design process should be considered. Sample design refers to what a sample consists of and how the sample is to be obtained. It is concerned with defining the population and frame, sample size, and sampling techniques.

A combination of sampling design and estimation method should be chosen so that the resulting estimates attain the best possible precision under the given budget, or to obtain the lowest possible cost for a fixed precision. The choice of sample design should take into account the availability of auxiliary information, as this can be used in the selection and the estimation process to obtain more accurate estimates. The sampling method for a survey can range from simple random sampling to a complex sampling and estimation procedure such as a multistage design. For more information on sample design see the SCH web site - Basic Survey Design Manual - Chapter 7.

Administrative collections by their nature collect data from all individuals in a certain population. Often, they do not involve the selection of a sample of units from the population and the derivation of statistics from such a sample.

However, where information is to be extracted from a very large administrative data set, it is sometimes effective to use a sample of the data set for statistical purposes. This can be for two reasons. First, the 'cleansing' of a very large data set can be a costly and time consuming task. Second, the analysis of very large data sets can also be time consuming and costly from a processing perspective. The resultant reduction of data may lead to only marginal reductions in accuracy, while substantially reducing the cost and time taken to produce results. In these cases good sample design principles must be followed in extracting the data. For longitudinal databases, methods that ensure a representative sample is maintained over time need to be applied.

## 3.5 QUESTIONNAIRE DEVELOPMENT

The main functions of questionnaires are to collect accurate and relevant information from respondents, record the data for processing, and provide a historical record. To achieve this effectively, a questionnaire should:

- clearly and concisely define what is to be collected and recorded
- enable respondents/interviewers to complete it accurately and within a reasonable time
- flow smoothly and logically from one question to the next
- use a language that is understood by the respondents
- avoid bias in question wording

- appear uncluttered
- provide suitable space for responses
- be easily processed by both people and systems.

More detailed information on questionnaire development can be obtained from the SCH web site - ABS Forms Design Standards Manual and SCH web site - Basic Survey Design Manual - Chapter 8.

For information on electronic data capture instruments, the following work, SCH web site - Towards Best Practice for Design of Electronic Data Capture Instruments, aims to develop standards and guidelines for electronic data capture instruments for national statistical collections from businesses.

## 3.6 SURVEY TESTING

Survey testing allows problems to be identified and corrected prior to the full survey being conducted. In some cases, data collected in the tests may be useful preliminary indicators of the survey results and can be used to estimate likely response rates as well as sample error, sample sizes and population variability. These preliminary results will involve a smaller sample and thus produce higher standard errors, which need to be taken into consideration.

Administrative systems should also be tested, including the process for turning administrative data into statistical information.

Survey testing provides guidance on the following aspects of the survey development process:

- adequacy of the sampling frame
- variability of the target population with regard to the subject which can be used in developing the sample design
- expected non-response rate and the effectiveness of measures aimed at reducing non-response
- appropriateness of the data collection or administrative method, including testing of various methods to determine the most suitable mechanism
- adequacy of the questionnaire or form, including testing of alternative versions to determine the most effective
- effectiveness of interviewer training and the adequacy of form instructions or robustness of data entry system

- answer categories to be used for pre-coded questions
- likely cost and duration of the survey or administrative system
- organisation of the survey or administrative system.

### 3.6.1 Types of Testing

The following six different types of testing are used at different stages of the survey's development and aim to test different aspects of the survey. Pilot tests and dress rehearsals are quantitative tests while the other types of tests are qualitative tests. For more information on survey testing see the SCH web site - Basic Survey Design Manual - Chapter 9.

**Focus groups**

Focus groups involve informal discussions of issues or topics with small groups of people from the survey populations.

They can be used early in the development of a survey to identify the language, concepts and possible definitions of terms in the survey. They should be conducted before a questionnaire is drafted, although they may also be used to test a few different possible question designs.

**Pretesting**

Pretesting is the process of informally testing questionnaire design with potential respondents.

The questionnaire is basically unstructured and is tested with a group of people who can provide feedback on issues such as the concepts, presumed level of knowledge for answering questions, range of likely answers to questions, how answers are formulated by respondents, obvious flaws and awkward wording of questions. The questionnaire should be redrafted after pretesting.

**Observational studies**

Observational studies involve getting respondents to complete the draft questionnaire in the presence of an observer. Whilst completing the form, respondents explain their understanding of the questions and the methods required in providing the information without interviewer intervention.

These can be useful for identifying problem questions, or the time taken to complete particular questions. Information on data availability and the most suitable person to complete the questionnaire may also be obtained.

### Pilot testing

Pilot testing involves formally testing a questionnaire or survey with a small representative sample of respondents. Semi-closed questions can be used to gather a range of likely responses which are later used to develop a more highly structured questionnaire with closed questions.

Pilot testing is used to identify any problems with the form or instructions to interviewers and interview times. It also allows the comparison of alternative versions of a questionnaire.

### Dress rehearsals

A dress rehearsal is a final trial run of the survey where the chosen sampling methodology is used to select a small sample from the target population.

Dress rehearsals are used to detect any problems that may arise in the survey design and/or processing system. They may also provide an opportunity to obtain data on survey costs and estimate population variances. Additionally, dress rehearsals should be able to provide preliminary information on the feasibility of the sample selection plan, fieldwork procedures and response rates.

### Post Enumeration Studies

A Post Enumeration Study (PES) is a study of a sample of respondents and non-respondents after a survey has been in the field, with the aim of evaluating the quality of the data. This can occur after a pilot test or after the final survey. It is usually done through structured, face-to-face interviews but can be done over the telephone, and utilises probing questions about how the respondent completed the form and their understanding of the concepts used in the survey.

Since the 1966 Population Census, each census has been followed by a PES, conducted by specially trained interviewers. The main purpose of the Census PES is to measure the extent of undercount and overcount in the Census. This is achieved by asking respondents if they were included on a census form for the household being interviewed, and if there were any other addresses where they may have been included in the Census. At each of these addresses (including the interview address), the personal information is matched to any corresponding census forms for these addresses to determine whether a person is counted, is counted more than once, or not counted at all.

The objectives of a PES can include:

- to evaluate the accuracy of data collected in a survey
- to test respondent understanding of concepts and definitions
- to obtain information on source data used
- to evaluate the design of the form and obtain information on respondent reaction to a new form
- to test the relevance of the data items included
- to gauge whether respondent load has been impacted on because of a new form
- to test whether changes to data item definitions and questions will result in significant changes to the data reported
- to evaluate whether an improved form design leads to more accurate and relevant responses from providers.

## 3.7 ADMINISTRATIVE DATA

The collection of statistical data by administrative systems is generally governed by program or policy issues. The data are obtained from agencies processing systems which have collected information as part of it's program responsibilities. The information is stored on databases which can be accessed electronically.

Fields, coding and edits can be added to administrative systems to serve a statistical purpose which is not an essential part of the administrative system. The addition of data items may improve the quality of output from the administrative system through cross checking and extra edits. However, any increased reporting burden on the respondent as a result of the extra data items should be minimised.

### 3.7.1 Extraction of Statistical Data from Administrative Systems

There are several different ways through which statistical data can be extracted from an administrative system. These include:

- Taking a snapshot (a picture of the database at a point in time) of the administrative database at regular intervals and loading it to a separate database system reserved for the statistical system.

This will ensure that statistical requirements, such as data aggregation or data item derivations, which differ from the requirements of the administrative collection can be built into the system from the beginning. For example, each quarter the NSW Police Service extract a snapshot of the crime recording system, the COPS database, containing all crime incidents recorded during the period and send it to the NSW Bureau of Crime Statistics and Research. The Bureau load it onto their own database for processing, including editing, and analysis. The Bureau publishes crime statistics annually on behalf of the Police. Using this technique police operations and their operational system are not affected by editing.

- Generating a separate statistical record during each administrative transaction, via a paper form or a computer system.

Specific statistical questions may be added to an administrative form. This has been undertaken on the Business Activity Statement completed for tax purposes and the Immigration and Customs forms completed on arrival to and departure from Australia.

Using a computer system, the transaction can automatically be added to the statistical system as well as the administrative system. This approach allows specific data items required by the statistical system to be defined and created simultaneously. This reduces the reprocessing required to transform the information into statistical data after extraction from the administrative system.

- An ad hoc approach to extracting statistical information from administrative systems is to query them as and when the need arises.

Although this may be practical in the short term it may not allow efficient processing, tracking and use of the statistical information.

### 3.7.2 Integrating Data from a Number of Administrative Systems

The collection of administrative data may involve the compiling of data from a number of different administrative sources. For example, data relating to individuals progressing through the justice system may be collected by different agencies within the government (e.g. police, legal aid, a variety of courts/tribunals, prisons). Each of these agencies will have administrative processes in place to collect the relevant data for their particular purpose, but there may be no mechanism for combining the data from these different administrative sources.

Agreement between agencies to standardise their administrative systems can facilitate the sharing of information, such as using consistent identifiers, for example the Australian Business Number. However, agencies should check the relevant confidentiality and privacy legislation before embarking on a data sharing exercise.

### 3.7.3 Policy Changes and Administrative Systems

In an administrative collection, policy and program considerations have a higher priority than statistical considerations in determining the concepts, definitions, coverage, frequency, timeliness and other attributes of the administrative program. As public policy changes and new legislation is introduced, program requirements and procedures will also change.

The data concepts and definitions used in administrative collections are closely linked with the programs being administered and will change as it changes. Changes can also be reflected in the survey population and the frequency of data collection in the administrative system. These changes can impact significantly on the statistical outputs produced. The ability to compare data over time is affected by changing concepts. These issues can be compensated for to some extent by documentation of concepts, population changes, etc.

Policy issues can affect administrative systems in many ways including the range of data that can be produced, the quality of the system and the uses to which the system can be put. Policy changes over time can also effect the consistency of a data set.

## 3.8 RESPONDENT LOAD

Respondent load is a measure of the effort, in terms of time and cost, generally measured in the time taken for respondents to provide satisfactory answers to a survey. Pilot tests or other forms of testing can be conducted to obtain an indication of the time taken by a respondent to complete the survey form.

Agencies should consider respondent load when planning any collection to ensure that it is justified by their objectives. Agencies should have policies and practices in place for managing relationships with respondents with the aim of keeping reporting load to the minimum practicable and maintaining cooperation to ensure the quality of collections.

Managing respondent load can be done by:

- consulting with users to ensure data collected is both relevant and meets their broad needs
- avoiding duplication with other collections
- using sound collection practices
- examining existing data collections before increasing the size of the collection due to a change in the users data requirements
- designing collections so reporting units can provide information easily
- selecting a collection method which suits the respondent, for example data should be collected in electronic form if it is already available in that form from respondent's records
- querying action arising from editing should minimise the number of follow-up contacts with respondents, in some cases computer-based imputation or estimation of missing data could be evaluated
- the Australian Government has a commitment to reducing the burden on business imposed by increasing statistical collections. Statistical collections affecting 50 or more businesses and run by, or on behalf of, Commonwealth Government departments and agencies are subject to a central clearance process. The purpose of clearance is to ensure that all such surveys were necessary and well designed to minimise respondent load and maximise benefit. For further details see the Statistical Clearing House web site www.sch.abs.gov.au.

**Chapter 4 - Data Collection and Data Extraction**

## OVERVIEW

Low response rates can lead to inaccurate estimates which results in incorrect conclusions being made about the population. Collecting the data in a professional manner can maximise your response rates and reduce errors.

Data collection has been split into two main components: managing respondents and managing the collection process.

Managing respondents provides advice on dealing with respondents in a professional manner in order to maximise response rates. Managing the collection process reviews the infrastructure that is required to convert the information provided by respondents into useable and reliable data.

## 4.1 MANAGING RESPONDENTS

Reliable data depends on public cooperation and goodwill to provide accurate and timely information as requested in data collections. Such cooperation and goodwill are maximised by assuring respondents that all data provided will be used for a recognised and community useful purpose and that individual confidentiality is maintained.

### 4.1.1 Initial Contact

Introductory letters advise respondents of their selection and their obligations. Often these letters will also explain the use of the information and the confidentiality of their responses. A contact name and number should also be provided in case the respondent has questions not covered in the introductory letter.

An introductory letter is an effective method of preparing respondents. Regardless of how you plan to collect the data, an introductory letter can still be sent to respondents. For example, if you are planning to collect data over the phone, a introductory letter could be sent out a week before calling. Introductory letters are often referred to as "warm contact" as they help to manage respondents expectations. SCH web site contains additional information about the use of introductory letters.

Privacy issues can also be covered in the introductory letter. In particular, linking administrative records to other sources can raise concern over the privacy of the data. Under the privacy acts, respondents should be informed that the data they are providing may be linked to other sources. Appendix 4 - Privacy and Confidentiality provides further information on the privacy acts.

### 4.1.2 Collecting the Information

Timing of data collection can impact on the quality of data and the response rate. There are times of the year where it will be difficult for respondents to complete your questions. During school holidays many families depart for holidays. Attempting to collect data in this period is likely to elicit a low response. Likewise, collecting information from business is likely to be more successful during business hours.

The questionnaire design can also influence response rates. A well designed form will have the following characteristics:

- short and simple questions
- a minimum number of questions
- easy to follow instructions
- uncluttered layout.

The principles of questionnaire design are explained in the SCH web site - ABS Forms Design Standards Manual.

In telephone or personal interviews, you may not be able to contact the respondent at the first attempt. A contact strategy should be in place to ensure that sufficient phone calls or visits are made to contact each respondent. Increasing the number of contacts will increase the response rates and decrease non-response errors. You may need to allow for up to 8 phone calls or visits to maximise your response rate.

### 4.1.3 Following Up

Additional contact with respondents may be required to ensure that data is provided within the required time frames or to clarify some of the data already provided.

In a mail based collection, several letters reminding respondents to complete the questionnaire may be required. Each letter should reiterate the purpose of the letter, state when the information was due and provide contact information (in case the respondent is having difficulty or has lost the form). Often these letters are sent two to three weeks apart.

After receiving the data, you may discover that there are missing values, surprising responses or logical errors. These types of errors should be detected by your processing system (refer to chapter 5). Some of these errors may be relatively minor, and do not need correction. However, larger errors may warrant recontacting the respondent to clarify whether the data your have recorded is a true reflection of their information. Many respondents see this additional contact as an intrusion as they have already participated in the collection. Due to this, these contacts should be minimised. SCH web site contains additional information about the use of follow up letters.

### 4.1.4 Dealing with Complaints

No matter how well your collection is planned, it is likely that you will find some respondents who are unhappy with certain aspects of your collection. You should allow for this by having a clearly articulated process for dealing with complaints. This can also included a review process in case the outcome of the complaint is regarded as unsatisfactory. All of you staff should be familiar with the complaints process.

## 4.2 MANAGING THE COLLECTION PROCESS

### 4.2.1 Training of Staff

The quality of training given to interviewers and processing staff has a strong influence on the end results obtained. As with any work environment, all of the staff working on the collection should have sufficient skills to fulfil their role. Staff who have direct contact with respondents may need to respond to queries about the collection, purpose or design. A list of frequently asked questions can assist these staff in handling these questions professionally.

General training of staff could include issues such as the purpose of the survey, the scope and coverage of the survey, an overview of the sampling methodology, the format of the questionnaire, interviewing techniques, recording of the responses, field practice, training in how to cope with queries and complaints, quality control and any administrative requirements.

### 4.2.2 Using Response Codes

As it is not realistic to expect your survey frame to keep up with all the changes in the real world as they happen, it is inevitable that the your frame may deviate from corresponding real world business characteristics. Some examples of the problems you may encounter are businesses which cease operating, people who are no longer relevant to your study or incorrect information on frame.

You may also encounter problems in the responses that you receive from the respondent. For example you may receive exceptionally large responses, only receive answers for part of the survey form or receive no answers at all.

Developing a simple coding system for responses will allow you to treat responses consistently, help you to manage your interactions with respondents and allow you to produce summary information. To develop a set of codes you should identify the basic problems that are likely to occur and assign a code for each problem. For example, you may decide to code a normal

response with no problems as 1, a business which has ceased operating as 2, a non-respondent as 3, etc.
Codes can be linked to your processes. In particular, your edits can be improved by relating the to these codes. For example a business which has ceased operating should not have any employees. In situations where the survey is repeated, codes can assist you in maintaining the frame. If a person has changed their name, you may wish to incorporate a code which tells your processing staff to make sure this is updated on the frame for future cycles.

### 4.2.3 Systems to Support the Collection Process

Despatch and collection control systems can vary significantly, depending on the collection method and available budget. These control systems can be as complex as an integrated purpose built system or as simple as a spreadsheet. Regardless of the system, the despatch and responses should be monitored through regular reporting of response codes. Regular monitoring can assist you in tracking of responses and refusals and prompt follow-up action.

Recent trends have seen despatch and collection control systems align with customer relationship systems. Customer relationship systems integrate your organisations contacts with each respondent, allowing you to develop a approach which draws on the respondents history.

## Chapter 5 - Data Processing

## OVERVIEW

Data processing refers to the process of turning collected data into an error free data file ready for the production of output. This chapter looks at data capture, coding, editing, validation, monitoring as well as weighting and estimation techniques.

## 5.1 DATA CAPTURE

Data capture is the process of transferring respondents information onto a computerised system.

Historically, data was captured from collection forms by physically keying it into a computer. However, the majority of data capture now occurs via an electronic process. Facilities such as optical character recognition (OCR), optical mark recognition (OMR), direct entry from Computer Assisted Telephone Interviewing (CATI) and Computer Assisted Personal Interviewing (CAPI), and electronic forms (or eforms) are now being utilised.

OCR and OMR scan respondents hand completed forms onto a data file for future processing while CATI and CAPI involve the interviewer entering data directly into a computer rather than onto a form. The advantage of CATI, CAPI and eforms is that responses are recorded in a data file as they are entered on the form. With electronic capture of the information, data entry occurs at the collection point which allows the collectors of the information to make queries at the source and alert the operator to possible errors in the data entry process.

## 5.2 CODING

Coding is the process of converting questionnaire information into numbers or symbols in order to facilitate subsequent data processing operations. For example, males may be assigned the value of one, and females two when coding an item such as sex. The code number can be printed on the form next to the response to assist data entry. Where a respondent answers a question in their own words (for example, 'What is your occupation?'), coding involves interpreting responses and classifying them into predetermined classes.

There are various forms of coding including manual coding, computer assisted coding and automatic coding. Manual coding involves entering a code from an index (i.e male equals 1, female equals 2). Computer assisted coding involves entering a truncated form of a response into a computer and then selecting from a restricted range of entries displayed on the screen with the code written automatically to the data file. With automatic coding, the computer system codes the information as data is captured.

The use of standard codes and classifications assists with comparing data both within the same collection and across different collections. Standard coding procedures and training materials are available for many standard classifications. More information about the use of standards is available from Appendix 3 Standards and Classifications.

## 5.3 EDITING

Editing aims to maintain data quality by verifying that values are appropriate, major errors are detected and outputs are consistent. The editing process can also assist in finding appropriate treatments to correct errors. By utilising edits, a number of non-sampling errors can be minimised. These include errors that were introduced during data entry or processing, recall error, interviewer error, non-response and invalid response. It can also quality measures about the collection and the process can suggest improvements for future surveys or administrative systems.

An edit is defined by specifying a test to be applied, the set of data it is applied to and the follow-up action to be taken if the test is failed. The test is a statement of something that is expected to be true for good data. Different edit types include validation/legality edits, missing data edits, logical edits and consistency edits (descriptions of these are below).

### 5.3.1 When does Editing Occur?

Editing can occur at different stages during data processing. Some of the stages of editing are:

- **Clerical editing**

  This stage includes all editing undertaken manually before the unit data are loaded into a computer file. Clerical editing in large collections is normally restricted to a visual scan of forms to ensure that important items of data are reported and that related items are completed.

- **Input editing**

  Input editing involves dealing with each response independently and is undertaken before aggregation of the individual estimates.

  In deigning an input edit system consideration should be given to tolerance levels, clerical scrutiny levels, resource costs, respondent load and timing implications. For example, tolerance levels for the invoking of edits should be set to avoid the generation of large numbers of edit failures.

- **Intermediate editing**

  Intermediate editing involves comparing related units or forms to ensure consistency between individual responses. For example, businesses in the same industry with a similar sized labour force would be expected to have similar responses to a survey on their output and costs of production.

- **Output editing**

  Output editing includes all edits applied to the data once it has been weighted and aggregated in preparation for publication. Output editing often focuses on identifying the units with the largest effect on collection outputs and ensuring that data for these units are correct and the consequent effect on outputs appropriate. For example, if a unit contributes a large amount to a subtotal or total, then the response for that unit should be confirmed.

There are usually a number of constraints (such as budget, time and number of people available for data editing) which will influence the design of the editing system.

### 5.3.2 Types of Edits

There are four main types of edits: validation edits, missing data edits, logical edits and consistency edits.

- **Validation Edits**

Check the validity of basic identification or classificatory items in unit data. Validation edits are input edits which are best done no later than data entry since failure of a validation edit often means further data entry is pointless until the error is fixed.

- **Missing Data Edits**

Check that data that should have been reported were in fact reported or that questions which should not have been answered weren't answered. An answer to one question may determine which other questions are to be answered and the editing system needs to ensure that the right sequence of questions has been answered.

- **Logical Edits**

Are used to ensure that two or more categorical items in a record do not have contradictory values. They can also be applied across time periods. For example, a respondent claiming to be 16 years old and receiving the age pension would clearly fail this type of edit.

- **Consistency (or Reconciliation) Edits**

Check arithmetic relationships between variables are obeyed. Determining these edits sometimes requires knowledge of the subject matter involved. Consistency edits could involve the checking of totals or applied across time periods.

For further information on editing see the SCH web site - Basic Survey Design Manual - Chapter 10.

## 5.4 DERIVATION AND VALIDATION

In some cases, data may have been collected in fragments. These fragments assist the respondent in completing the form. For example, if you are attempting to collect total household income you may ask each person in the household to provide you with their individual income. The information from each person would then need to be compiled to derive the desired variable. This process is referred to as derivation.

Validation is the scrutiny of survey or administrative data against other sources of information. These other information sources should generally tell a story that is consistent with the survey data. It is an important aspect of ensuring data quality.

Generally, validation is not a prescriptive process. Some examples of validation methods are

- comparing output with complementary data sources such as other statistical and administrative collections in the same subject area
- comparing trends in data with comment by experts in media or public fora
- comparing management information produced by processing systems, particularly where major differences between collection cycles are noted
- looking for odd trends between variables
- using a subject matter expert to examine the data to see if it looks right.

## 5.5 MONITORING

### 5.5.1 Monitoring Performance

The data processing cycle should be monitored in order to keep the collection manager and operational staff informed of:

- the progress of forms through the processing cycle;
- the resources being used by the system (time and cost); and
- the reasons for (and the magnitude of) the changes made to the data.

For example, information about the above aspects of the system can be used to monitor deadlines and to see whether resources need to be shifted from one facet of processing to another to ensure that the work continues on schedule.

Performance information will help decide if processing has been cost effective or whether some aspects of the editing system need modification in order to produce a more efficient and effective editing system for the next time the collection is run. An editing system also needs careful reviewing after modifications have occurred to see if the changes have been made correctly and whether they have proven beneficial.

### 5.5.2 Monitoring Quality

System wide quality monitoring information assists in examining what effect processing has had on data quality and to determine whether existing edits are sufficient. The monitoring should look at the number of incomplete records, the number of units that failed edits, the changes in data quality as the data passes through the system, the effect of edits on final estimates and any problems with the questionnaire. A high edit failure rate could indicate poor questionnaire design or tolerance levels in the editing system that are too strict.

Some of the procedures that can be used to monitor the quality of the processing system are listed below:

- Random auditing where a sample of responses is audited. For example, some collections contact a sample of respondents to confirm the validity of the original responses.

- System management reports produced automatically at different places or stages in the system. For example, reports could be produced on the number of responses at different points in the processing system or provide a summary of records amended at given points of time. These can then be compared to records kept for previous periods.

- Reports summarising variable information. These can be used to check whether data are out of range or inconsistent and whether data have been sufficiently edited.

- Data snapshots can be taken to enable comparisons of outputs. This will ensure that the data can be checked for errors and standards consistently applied. The snapshot time will depend on the level of analysis undertaken on the data. The creation of a snapshot file of the data allows for quality checks and editing of the data to be undertaken. These data can be used to respond to data requests and ensures that later analysis remains comparable.

- A register of problems (and their solutions) can be maintained. This may then be used as an ongoing guide for the resolution of recurring problems, and is also a valuable tool at the time of evaluation for identifying and improving ineffective processes.

## 5.6 WEIGHTING AND ESTIMATION

Sample surveys only take information from a subset of the population. The process where information from a sample is expanded to represent the entire population is called estimation.

To create estimates which represent the entire population, weights are required. Weights are a multiplier which is applied to the information from the sample. The methods for calculating the weights will vary and depend heavily on the design that was used to collect the sample.

Measures of the accuracy of these estimates will assist you in understanding the quality of the data. There are a range of measures such as variance, standard errors or relative standard errors. For surveys which collect a large amount of data, it may not be feasible to create measures of accuracy for every data item. In this situation, you should consider creating measures for a smaller representative set of data items.

More information on weighting and estimation is available from the SCH web site - Basic Survey Design Manual, Chapter 11. For complex surveys weighting can be extremely complex, and it is often useful to consult a statistician.

# Chapter 6 - Statistical Analysis and Interpretation

## OVERVIEW

Analysing and interpreting data assists in identifying important characteristics of a population, providing insights into the topic being researched. Techniques used can range from simple summary statistics (e.g. mean and median) to quite complex analysis (e.g. econometric modelling and statistical regression). This chapter examines the major issues to be considered when analysing results from a statistical collection.

## 6.1 STEPS IN ANALYSING DATA

There are many different approaches that can be made to analysing data. One method, referred to as the research question approach, consists of the following steps:

1. Identify issues and formulate questions to be addressed by the research.
2. Develop initial models or hypotheses that will be tested during the analysis.
3. Obtain appropriate data (from an administrative data source or survey collection).
4. Analyse the data (i.e test hypotheses, relationships and models).
5. Quality assure results taking into consideration any constraints in the data.
6. Present results, including any assumptions and constraints.

These steps are discussed in more detail below.

## 6.2 RESEARCH ISSUES AND OBJECTIVES

The research objectives of the project should be stated in the planning stages. It should cover key policy issues such as monitoring the effectiveness of a particular program through to the economic and demographic behaviour of a particular population.

Once data is gathered, it may become apparent that the research objectives cannot be fully answered. If this is the case the research objectives may need to be reviewed and possibly narrowed. For example, a particular section of the target audience may not have fully responded to the survey, or respondents may not have the necessary information to complete a particular question or questions. Any limitations in the data need to be taken into account during the analysis phase (see Section 6.5 - Analysing the Data).

## 6.3 DEVELOP MODELS AND HYPOTHESES

Preliminary models and hypotheses should be developed from analysing the relationships between the variables that are to be collected. These data models should be used in the initial testing in the data analysis phase (see Section 6.5 - Analysing the Data).

## 6.4 USING APPROPRIATE DATA

By understanding the objectives of the research the analyst will be able to select the most relevant data and apply the most appropriate statistical techniques. As part of selecting the data, the analyst needs to assess and document it's quality. The quality of statistical information is the degree to which the information describes the items of interest. If the survey was not designed to measure the items of interest, then it may not be appropriate to use the survey results for the proposed analysis. More information on how to assess the quality of statistical information can be found in Appendix 7 - Quality Declaration and Assessment.

Before starting to analyse a data set, the analyst should become familiar with the data (i.e reviewing the documentation, ensuring that the values and ranges of variables are in line with expectations). Although most data has already undergone editing in the processing stage, additional edits may be required to ensure the data is appropriate for the intended use.

Any statistical analysis should be documented so that others are able to access the analysis performed. A copy of the original data, documentation which explains each dataset and how they are used in the analysis, and clearly labelled programs used to manipulate the data should all be kept and made easily accessible.

## 6.5 ANALYSING THE DATA

Many people find it difficult to interpret bare numbers. The process of turning data into information can be thought of as the conversion of numbers into text. Descriptive analysis such as means, measures of dispersion and percentiles are very useful as it can highlight aspects that the reader might overlook. Summary measures also allow data to be compared with other data, or allow aspects of the data (e.g. characteristics of population groups) to be compared.

Any statistical analysis should focus on the research objectives. It enables testing of hypotheses about the attributes of the data and complex investigation into relationships in the data. Modelling techniques such as linear regression, logistic regression, and survival analysis are one way to explore these relationships. Assistance should always be sought from an experienced analysis when conducting complex analysts.

Statistical methods for analysing data can be characterised according to the type of data to which they are applied. The field of survey statistics usually deals with cross-sectional data describing many different individuals or units at a single point in time. Time series data describing a single entity across time. Longitudinal data blends characteristics of both cross-sectional and time series data. Like cross-sectional data, it describes each of a number of individuals and like time series data, it describes each single individual through time.

For more information on statistical analysis techniques and the statistical methods used to summarise results, show different types of distributions, make estimations and identify outliers see the SCH web site - Basic Survey Design Manual - Chapter 11. The ABS Information Paper, An Introductory Course on Time Series Analysis, May 2000 (cat. no.1346.0.55.001) introduces some of the theory underlying time series analysis, discusses different seasonal adjustment philosophies and looks at some of the major issues relating to seasonal adjustment .

## 6.6 PRESENTING A COHERENT STORY

The presentation of results should be tailored to address the aims and objectives of the survey and to satisfy the potential users of the results. Statistical outputs are often best expressed by having the main results of the data reported as tables, with graphs and commentary to facilitate understanding. The report should convey the main features clearly and follow a logical progression, use as little jargon as possible and provide clear insight into the data. It should include information on how the data were collected, compiled, processed, edited and validated and it should point out any known limitations.

As a general rule the report should include the following:

- Executive summary (summarising the main objectives and findings)
- Introduction (setting out the purpose and aims of the survey, background to research, defines terms and concepts etc)
- Methodology (describes method of sampling and information on survey population, data analysis and statistical procedures used)
- Findings and analysis (details of sample numbers, response rates, results and interpretation of tabulations)
- Conclusions and recommendations (summarising major findings and outline future actions)
- Appendices and references (i.e copy of the questionnaire).

### 6.6.1 Statistical Presentation

Presenting data graphically can often highlight the main points of interest and relationships between variables, indicating which analyses should be considered.

Although graphs are generally easier to interpret compared to tables, they can also hide information. Therefore analysis and presentation of data should not be restricted to graphs. There are many different types of graphs and the decision as to which one to use will depend on the type of data to be presented. For more information see the SCH web site, Basic Survey Design Manual, Chapter 12 for details on statistical presentation for graphs.

**Chapter 7 - Dissemination of Statistics**

## OVERVIEW

A dissemination strategy for delivering information to users will assist in maximising the usefulness of survey or administrative data. This chapter examines aspects to be considered in a dissemination policy including timetabling, methods of release, timeliness of release, access to data, confidentiality protection, data quality assessments and metadata. It also discusses the related issues of revisions and corrigenda, requests for information and misleading interpretations.

## 7.1 DISSEMINATION PLAN

A data dissemination plan should outline how the organisation or agency intends to release the available information. In developing a plan any legislative obligations, demand for the information, dissemination processes and any other factors relevant to the dissemination of the product should be taken into consideration. Catalogues and directories should be made available so that potential users are aware of what statistics are available and the expected release dates. The dissemination plan should ensure wide access to the data.

Within the dissemination plan a pricing policy may need to be developed, taking into consideration the following issues:

- public good requirements or community service obligations

- whether the organisation will charge for the information and the basis for such charging, for example:

Australian Government agencies should follow the Commonwealth Cost Recovery Guidelines for Information Agencies.

## 7.2 TIMETABLE FOR RELEASE OF STATISTICS

Timetables should plan to release statistics as soon as possible after their collection. However, they should also be realistic and achievable. Elements to consider in developing the timetable include:

- release date should be specified in advance

- approval to release data
- checking and validation of data

- documentation on data items and how to interpret the data.

## 7.3 TYPES OF DISSEMINATION

Data may be released under a number of formats. These comprise:

- printed publications;

- electronic publications (including Acrobat PDF file format);

- core data standard products, which are products prepared in anticipation of general user demand or for a substantial number of users (in various media);

- services which provide access to "packages" of statistics, for example, AusStats which provides subscribers with on line access to all ABS publications and a range of other data;

- newsletters or short articles;

- microfiche;

- telephone or facsimile responses to special data requests;

- customised products and services which are made to order for specific customers;

- aggregate data files; or

- unit record files available to authorised users through data centres and data laboratories or for public use. (Any such release would need to be consistent with the agency's legislation and confidentiality and privacy principles).

The information may be released using a combination of different media, with some products meeting general needs and other products catering for more complex needs. The statistics should be released in forms convenient to users.

## 7.4 TIMELINESS

The timeliness of statistical information refers to the delay between the end of the reference period to which the information pertains and the date on which the information becomes available. If a range of statistical outputs is to be released, broad or preliminary statistics can be released first, followed by more detailed releases based on more complete data and analyses. Data can be also be released with qualifying statements regarding quality and the amount of editing that has been done to the data if timing is an issue.

## 7.5 OPEN/EQUAL ACCESS

Open access to official statistics on the economic and social condition of a country and its population is an essential element of a democracy. A dissemination policy should ensure that all users, including the general community and government, has access to important statistics relating to government programs and activities. To ensure this, it may be necessary to release information using more than one of the formats set out in Section 7.3 Types of Dissemination above, eg releasing information electronically and in printed publications to cater for users who do not have access to a computer.

However, many administrative collections are undertaken for a very specific purpose and consequently the release of statistical outputs to individuals or groups needs to be carefully considered. Some administrative collections are authorised by legislation and specifically preclude the release of certain information. Adherence to such legislative or government directives must always be followed, but this should not discourage the widest permissible dissemination of statistical outputs to users.

## 7.6 CONFIDENTIALITY

Data confidentiality i.e. assurance that information about individual respondents is not released outside the organisation and cannot be derived from published results, should be supported by the dissemination process. See Appendix 4 - Confidentiality and Privacy for further information.

## 7.7 DATA QUALITY SUMMARY

The aim of a data quality summary for disseminated data is to explain the quality and other aspects of the data, so that users can make informed and appropriate use of the statistics. Users need to be informed about the survey design, processing, methodology, collection and any other factors that might affect the magnitude of error.

A Quality Declaration template (see Appendix 7 - Quality Declaration and Assessment) to assist agencies in assessing their data needs. The information contained in a Quality Declaration can be adapted to create a brief data summary noting various issues relating to data quality e.g. relevance, accuracy, timeliness and consistency, to accompany the disseminated data.

## 7.8 METADATA FOR DISSEMINATION

Metadata is textual information about the actual data. It is important that metadata is provided with disseminated data to allow users to properly understand and use the data. This metadata should describe the output datasets fully and consistently.

All agencies should provide metadata holdings on their web site. This makes visible and relatively easy to access.

For example, the Australian Institute of Health and Welfare's "Knowledgebase" can be accessed via the AIHW web site. The Knowledgebase is a metadata registry for Australia's health, community services and housing. This electronic storage site provides open access to view and comment on data definitions and standards related to Australian health, community services and housing assistance.

For further information about metadata, refer to Appendix 2 - Data Management.

## 7.9 REVISIONS AND CORRIGENDA

The release of revisions to data should balance the need for users to have the best estimates, with the uncertainty created by frequent revisions. A revisions policy should be made available in the explanatory material associated with the data, especially where the estimates are preliminary or contain seasonally adjusted, trend or constant price series. A corrigendum draws attention to errors and omissions in a released publication and provides the corrected or omitted information.

## 7.10 REQUESTS FOR INFORMATION

Agencies should have a system in place for servicing client requests for available data so that users' information needs are met and improved decision making within governments and the community is encouraged. Important issues to be considered as part of a client request service include:

- confidentiality of data to be released
- pricing policy is uniform across the organisation
- that there is ownership of the request (it may be desirable to prepare pro forma responses to frequently asked questions)
- that there are systems in place to provide contact detail on all publications, electronic and hardcopy, to ensure that requests and queries go straight to the most appropriate person.

## 7.11 MISLEADING INTERPRETATIONS

Agencies are entitled to comment on erroneous interpretation and misuse of their statistics. For example, an agency may comment when there is a serious error of fact or when the comment impinges on the reputation/credibility of the agency and its statistics.

The response to a misleading interpretation of statistics is most commonly a letter to the newspaper or journal that reported the matter. It could also involve a detailed set of questions and answers being developed to respond to media enquiries which can form the basis of a public relations strategy, changing the presentation of the statistics and educating users through analytical articles or specific training.

**Example**

Youth unemployment is an issue of considerable political and community concern, involving much debate on possible solutions. The Australian Bureau of Statistics provides a range of statistics on youth unemployment in its labour force publications. However, in policy debates and media commentary concerning youth unemployment, these statistics are often misinterpreted or misused. To assist in correct interpretation of the data, the Australian Bureau of Statistics is:

- enhancing the presentation of teenage labour force and education data from the monthly Labour Force Survey
- introducing supplementary measures
- reviewing articles on youth unemployment published annually in Labour Force, Australia, to ensure that they assist in the education of users of these statistics

# Chapter 8 - Evaluating the Statistical Activity

## OVERVIEW

Evaluation is a key component of the development cycle of any statistical activity. It is a measure of how well the statistical activity has met the objectives of the collection. Evaluation provides the necessary information to improve the procedures and content of the statistical activity. Ideally, a statistical activity will be evaluated at each stage of the statistical cycle and then be also subject to a final evaluation. This chapter examines various evaluation strategies and the different types of formal statistical reviews that can be undertaken.

## 8.1 EVALUATION STRATEGIES

Evaluation procedures should be implemented as part of an ongoing improvement process for agencies involved in data collection, processing or dissemination. The evaluation strategy becomes as much a part of developing any new statistical activity as they are in reviewing an old one.

An evaluation strategy should specify a standard set of quality measures and key performance indicators (quantitative and qualitative), and how these will be measured or assessed. In agencies where a statistical activity is shared across different areas, it is useful to specify reporting responsibilities to the evaluation process.

The choice between which evaluation techniques to use will depend on which stage of the statistical activity is being evaluated. Key areas to include in an evaluation strategy might be:

- standard quality measures of data collections - these might be included in response to questions arising from a quality declaration format Appendix 7 - Quality Declaration and Assessment
- a measure of outcomes against the initial project specifications
- process mapping with key performance indicators
- an assessment of clients use of data against survey objectives
- a review of client satisfaction with products
- a project management assessment - monitoring the achieved timetable versus planned target deadline
- a cost assessment measured against the proposed budget
- an evaluation of the field performance of a data collection

The extent of the evaluation should match the complexity and size of the collection or data purchase. Many of the questions suggested in Appendix 7 - Quality Declaration and Assessment should be asked by the agency or program as part of an evaluation of a collection or dataset. Using this framework to answer key questions about the quality of the collection will assist in developing further improvements in content, procedures and processes. This is particularly relevant when the activity is to be repeated. The results of the evaluation should be well-documented and stored along with the other documentation pertaining to the collection or dataset.

## 8.2 FORMAL EVALUATIONS OF STATISTICAL PROCESSES, SYSTEMS AND SURVEYS (REVIEWS)

Formal evaluation procedures within agencies are often known as reviews. These reviews may be conducted periodically or in response to significant quality issue concerns. Periodic reviews are a beneficial means to maintain and improve ongoing administrative and statistical systems to ensure that a high level of quality is upheld by the system and that the objectives of collections are meeting user needs. Reviews may also be prompted when significant changes to a statistical process have taken place, for example, the redevelopment of a survey's computer system, a changed budget, or known changes in user needs. These reviews aim to assess the impact of significant changes on statistical activities.

A review is a systematic evaluation of statistical activities to assess whether users needs are being met, whether the methods used and the quality of outputs produced are appropriate, and whether more timely or efficient options are available. It is beneficial when conducting ongoing statistical activities or prior to conducting a new activity, that the process, objectives and systems be subject to an independent review. This provides a mechanism to highlight potential difficulties and improvements in process or methodology that may not be apparent when developing a statistical collection. This review may be conducted by other areas within the agency or from outside consultation.

### 8.2.1 Types of Reviews

Reviews may fall into four broad categories which are outlined below. These are effectiveness, efficiency, quality and other alternative reviews.

**Effectiveness reviews**

These reviews examine how well the outputs of have satisfied the stated objectives of the activity. The types of effectiveness reviews undertaken include:

- adequacy (fitness for purpose) of statistical products or services. These reviews measure the outputs of the statistical activity set to preset standards and criteria;

- appropriateness of statistics/work program. This appropriateness is a measure of relative priority and highlights the potential for improving the type and number of outputs from a particular work program;

- who is using the statistics and how - this type of review is targeted at whether the survey outputs are meeting key user needs and how these outputs are best delivered (what medium is being used e.g. electronic, paper); and

- new statistics (what is required to meet policy initiatives e.g. regional data, rural data).

**Efficiency reviews**

In general, efficiency reviews will evaluate whether there is a reduced cost method of achieving the survey objectives. An efficiency review may uncover inefficiencies in the production of statistics such as duplication of effort, non-priority work being undertaken, or inappropriate information technology operations. Efficiency reviews may also highlight alternative sources of data, reducing survey costs and respondent load.

**Quality reviews**

These reviews involving applying quality measures to the delivery of outputs. The quality measures outlined in Chapter 3 - Designing the Statistical Activity may act as key indicators in the review process to identify areas of possible quality improvement. Quality reviews include:

- evaluating key quality attributes of the statistical activity. These cover relevance, accuracy, timeliness, accessibility, coherence and interpretability
- examining client satisfaction
- reviewing methodology
- examining standards and classifications
- examining and reviewing codes of good practice
- systems which support collections
- improving service level standards.

To measure the cost/benefit of data systems, quality indicators should be developed for inputs (costs) and outcomes (benefits). This would require a mix of quantitative and qualitative measures, particularly for outcome indicators. Quantitative measures for inputs could be derived from techniques used in activity based costing. When establishing a quality based assessment framework, it is important to consider not only quality indicators but also the collection of information and procedures for monitoring, reporting and providing feedback on quality results.

## Appendix 1 - Phases of the Statistical Cycle

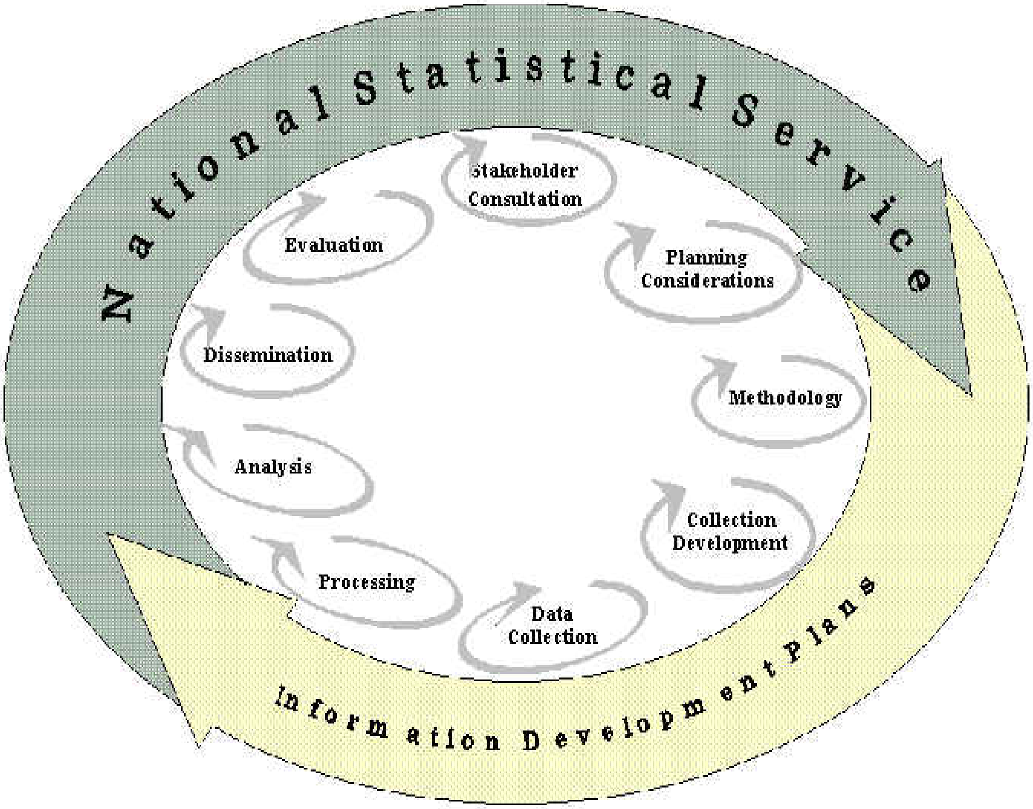


**Diagram 1 - Phases of the Statistical Cycle**

As illustrated above, the statistical cycle can be broken down into nine phases of activity. These are further defined below.

While the activities are shown in a chronological order, some phases may be undertaken concurrently and activities such as client liaison and planning may be undertaken throughout the cycle.

**Stakeholder consultation** involves determining who are the agency's/project's stakeholders, what are their/our needs, and how can these needs be met effectively. Stakeholder consultation also involves discussing collection objectives and content, data priorities, and output requirements and products.

In **planning considerations** decisions need to be made regarding the objectives of the collection, budget and resource constraints, and sensitivities. Research into the subject matter, statistical process, and existing data is generally conducted prior to developing a new statistical collection. This phase also involves determining legislative requirements and constraints, examining other relevant collections, and establishing a project management plan.

The **methodology** phase involves determining definitions, classifications and standards, and scope and coverage of the collection. This phase will also include examining whether to run a census or sample survey or to access existing data sources for the data collection.

**Collection development** is concerned with developing the frame, sample design and methodology. It also involves determining the most appropriate data collection method, and then spending time on questionnaire design as this influences the accuracy of survey results. Survey testing at this stage can provide guidance on many aspects of the survey design. Opportunities to reduce respondent load can be considered. Making data specifications and identifying output requirements are also part of developing the collection.

**Data collection** involves managing the collection phase with an appropriate dispatch and collection control system, maintaining collection documentation, providing interviewer training and instructions for face-to-face or telephone collections, monitoring response rates and the quality of the data collection process.

**Data processing** aims to produce a file that is free from errors and that can be manipulated to produce statistics. This generally involves coding, checking, data entry, editing, monitoring the response rate and monitoring the whole processing procedure. Privacy and confidentiality of information must be taken into consideration.

**Statistical analysis** can take place once estimates have been produced. This analysis can range from simple tables and graphs to more complex analytical techniques such as time series analysis. Analysis requirements should be factored into the planning stages of the survey.

**Dissemination** involves determining the most appropriate methods of delivery of the statistics to users. Statistical information may be disseminated both electronically and through publications and other products. The data may need to be confidentialised prior to release to ensure that tables and other products do not release confidential information about any individual respondent or organisation. A pricing policy may need to be followed to ensure consistent pricing across the organisation.

**Evaluation** encompasses monitoring the quality of outputs (in terms of accuracy, timeliness, relevance, accessibility, interpretability and coherence), and devising strategies for improving these quality measures. It is also necessary to monitor and evaluate processes (e.g. editing, data entry) to determine the causes of errors and delays, and to take action to improve performance. In addition, the objectives and procedures should be evaluated and modified. Evaluation comprises qualitative (e.g. Post Enumeration Surveys) and quantitative (e.g. survey response rates) approaches.

## Appendix 2 - Data Management

Metadata
Additional Documentation
Data Directory/Catalogue/Dictionary
Data Storage
Data Warehousing
Data Retention

Data management is the handling of data from acquisition and input through processing, output and storage. The objective of data management is to ensure that data holdings are of a sufficient level of quality to meet user needs. Data managed effectively is:

- **Relevant** to the information needs of users
- Easily **accessible** to users

- Documentation allows users to **interpret** whether the data is suitable for a specified purpose
- Data is **coherent** with other data sources.

The concept of metadata is central to data management. This concept is expanded upon in the Section below on metadata but it is information that describes data. Metadata enables users to assess, independently of the producer, the suitability of data relative to their needs.

Data are more readily managed efficiently and effectively if the appropriate systems and methods have been set up for the cataloguing and storing of datasets. There are many different methods available to manage data holdings and a warehouse is one method that can be considered.

## METADATA

Metadata is information which describes the actual data. Metadata assists data users to decide whether a particular data set is fit for a given purpose. It can also be used by people managing data as a guide or reference manual, or to evaluate system or procedural performance.

The National Statistical Service initiative will be recommending a set of minimum metadata standards for government statistics. A list of possible metadata items is listed below.

**Suggested Elements of Metadata**

Administration

- name of organisational unit responsible for data source
- contact officer (include position title and section)
- more detailed information about data source (e.g. URL of web site which contains relevant information).

Name and Overview

- name of the data source
- frequency of data source
- scope of data source
- statistical unit(s) (e.g. persons, households, businesses)
- the type of information produced
- main breakdowns (e.g. geographic).

Purpose

- purpose of the collection (i.e. the specific policies, issues, or actions that are being determined or assessed).
- key data outputs
- main users/uses of the data.

Scope and Coverage

- the scope (or target population) is the set of units about which information is required (e.g. residents of private dwellings in all areas of Australia)
- the coverage refers to the actual set of units about which information can be obtained or inferred.

Conceptual framework

- frameworks describe the concepts associated with a topic and organise them into a logical structure. Frameworks also show the key relationships, processes or flows that exist between elements.

Accuracy

- sources of error e.g. processing error, coding error

- if survey, sample size, percentage of population sampled, response rate, and sampling error

- areas where careful interpretation is required.

Main outputs

- units of collection e.g. persons, retail establishments, transport etc
- data breakdowns - levels at which the data are to be disaggregated e.g. by state, by industry, by sex etc

- type of statistics, i.e. level or movement, totals, means, medians, proportions or indexes, seasonally adjusted or trend, constant or current prices.

Classifications

- classifications used by the collection including the minimum level at which estimates are calculated and published.

Other concepts

- definitions of other key terms and concepts relating to the data source.

Geographic detail

- geographic areas for which data are available.

Frequency

- the time interval between collection cycles (e.g. once only, monthly, quarterly, annually, 5 yearly)

Data source history/ coherence

- the history of the data source, including the data available over time and key milestones in releasing data.
- a list of past major changes to the data source that impact on outputs and comparability over time (e.g. changes in target population, frequency, content, collection methods, sample design, benchmarking, etc).
- comparability with other sources.

Data available for dissemination

- frequency of data release
- form of data release, i.e paper publication, internet
- the time period between collection and release (to provide an indication of the timeliness of the data).

## ADDITIONAL DOCUMENTATION

In addition to metadata, more general documentation relating to various aspects of data collections should be kept and stored for future reference. For example, documentation covering the reasons why certain decisions were made, minutes from meetings, project proposals, changes to collections, etc can serve as a record for users and producers to ensure that data are interpreted and used correctly.

## DATA DIRECTORY/ CATALOGUE / DICTIONARY

In addition to being well documented, data should be easy to search and locate. Recording data collections and databases in a directory, catalogue or dictionary makes the retrieval of information easier and promotes the transparency of agency data holdings. The directory should contain facilities for searching by topic or theme, geographic area and keywords.

## DATA STORAGE

Data storage systems are used as repositories for the data collected in statistical or administrative collections. For simple collections data may be held in a spreadsheet, while more complex data holdings will be kept in a database or data warehouse.

Listed below are some of the major issues to be considered in designing a data storage system:

- access controls to the storage systems
- adequate documentation on systems and procedures
- the use of standard classifications across an agency's collections
- appropriate backup and retrieval policies
- appropriate data retention and archiving policies which allow for archiving of files no longer in regular use
- management of disk space used by collections
- availability and accessibility of metadata
- naming conventions and cataloguing for identification of file contents.

## DATA WAREHOUSING

A data warehouse is comprised of data brought together and specifically structured for querying and reporting. The main output from data warehouse systems are either tabular listings with minimal formatting or highly formatted 'formal' reports. A data warehouse can be a relational database, multidimensional database, flat file, hierarchical database, object database, etc.

There are a number of benefits in having data stored as part of a data warehouse. For users, it can mean that a comprehensive catalogue of data holdings is available. A warehouse can provide a system with facilities for manipulation, extraction and dissemination of statistics from stored data together with the metadata relating to their definition, collection and processing.

For the producing agency, a data warehouse can make it easier to maintain standards for storage, access, use and dissemination of the data. It also provides a central vehicle for the storage of concepts, definitions and procedures for agencies collections. A warehouse brings efficiencies to data storage by supporting the sharing and re-use of metadata across different collections.

## DATA RETENTION

The underlying justification for the retention of any information is the presumption that at a future date the information is likely to be accessed to satisfy internal and/or external user demand.

The objectives of a data retention policy are to ensure that:

- decisions to retain data are based on an assessment of the future need for that data
- retention is justified for a specified period or until a specific event occurs
- the optimum data storage medium or combination of media is selected with regard to overall costs and user requirements
- all collections have documentation covering their data retention practices.

The retention and/or destruction of Commonwealth records is governed by the National Archives of Australia (Archives) under the "Archives Act 1983".

Section 24 of that Act provides that agencies can only dispose of records (i.e. destroy, transfer custody or ownership, or damage or alter):

- if required by another law
- with the approval of the National Archives of Australia

- as a normal administrative practice (i.e. routine procedures that dispose of records of only ephemeral value - duplicates, telephone messages, compliments slips etc)
- for the purpose of returning records to the custody of the Commonwealth.

National Archives of Australia approves the disposal of records by issuing disposal authorities. These are legal instruments that describe groups of records and state the minimum periods that they should be kept. The Administrative Functions Disposal Authority was developed by National Archives of Australia to cover records that are common to all Commonwealth Agencies. Specific Disposal Authorities for an agency's operational records are developed by agencies in consultation with National Archives of Australia.

Further information on the archives policies of Australian and State/Territory Governments can be found at:

National Archives of Australia
State Records Authority of New South Wales
Public Record Office Victoria
Queensland State Archives
State Records of South Australia
State Records Office of Western Australia
Archives Office of Tasmania
Northern Territory Archives Service
ACT -Territory Records Office

**Appendix 3 - Standards and Classifications**

Standards
Classifications
Integration

ABS Standards and Classifications
International Standards and Classifications

The use of a comprehensive set of statistical and methodological standards allows an integrated and meaningful statistical picture to be provided. It makes it possible to draw all the data from different sources and at different times about a particular topic, variable, or population, together in a meaningful way. This allows comparisons and links to be made between and within datasets e.g. time series analysis, and integrated frameworks to be developed e.g. different types of health frameworks - those which endeavour to identify the relationships between various health related factors, and those which primarily provide a reporting structure. The purposes of this integration are to improve the usefulness, reliability and comprehensiveness of the data and to reduce duplication of data collection.

**Statistical standards apply to:**

- data item definitions
- concepts
- statistical units
- classifications
- coding processes
- derivation procedures
- question modules

**Methodological standards apply to:**

- sample design issues
- conceptual issues
- collection methodology issues
- processing issues
- data standards
- dissemination

This appendix covers the use of standards and classifications, integrating data and provides examples of standards and classifications.

## STANDARDS

Standards for statistics assist in maximising the effectiveness of statistical outputs and the efficiency of the production process. That is, effectiveness in terms of support for comparability (over time, space, industry, etc) and coherence (i.e. the capacity for integration) of the statistics. The use of statistical standards permits collection of statistics on a consistent basis over time.

While comparability and coherence are important for any dataset, they are particularly important where data are obtained from multiple sources and have to be combined in some way, or where the outputs are used in a wide variety of contexts.

For example, the comparison of data may be aided by the use of standard collection units. Collection units are the units of observation e.g. businesses, farms, motor vehicles, building sites, persons, households and families.

Statistical standards relating to the statistical process can be as important as those related to the statistical product. Standards may reduce the resource requirements associated with many aspects of survey development and maintenance.

## CLASSIFICATIONS

Classifications for statistics facilitate the accurate and systematic arrangement of data into categories according to common properties. The use of standard classifications result in statistics that are consistent and comparable over time and across different sources. Classifications aim to achieve economy of effort in developing quality information.

Comparison across different geographical regions is an important focus for many administrative datasets and statistical collections. Using standard classifications assists in this process. Administrative data are an important potential source of small area data which can be combined with other data sources to provide agencies with a more complete picture of the population of interest.

Statistical classifications can be used for aggregating and disaggregating data sets meaningfully for complex analysis, including the construction of indexes. They can also be used to construct a classification for a different variable on the basis of the classifications for two or more component variables e.g. socio-economic status classifications typically have categories defined by reference to categories found in classifications for occupation, labour force status, and educational attainment.

## INTEGRATION

The integration of data allows data to be used beyond the immediate purpose for which it was produced. The potential benefits of statistical integration include:

- more coherent data - statistics from different collections can be compared through the use of common data items, classifications, and other terminology.

- more efficient systems - use of common components may lead to less duplication.

- provider load is reduced - less duplication of data items collected. For example, an organisation might be supplying employment data for three different surveys, but if the standard definition for employment was used, the data could be supplied only once and the different data requestors could utilise the same information.

Frameworks exist for integrating and presenting data in many fields. The use of standards and classifications in frameworks greatly reduces the effort in integration and reconciliation required when bringing data together. Integrated economic statistics involve collections of economic statistics on the basis of an overall framework, for example the frameworks detailed in the System of National Accounts and the Balance of Payments Manual.

### Example

The Australian Bureau of Statistics has developed a framework for Integrated Economic Statistics under which most of its business surveys are conducted. This framework requires that the statistical structure of each business entity is based on a standard units model, and an industry code is allocated based on predominant activity. A key component of the framework is a centralised Business Register which stores this information about each business and which is used to produce population frames for collections of economic statistics, based on the industry code. A set of standard classifications is used to describe characteristics of businesses, such as their industry, number of employees, and type of legal organisation. The classification used is the Standard Institutional Sector Classification of Australia (SISCA). In addition, standard data item definitions have been developed, and these are accompanied by the use of standard questions in questionnaires.

## AUSTRALIAN STANDARDS AND CLASSIFICATIONS

Statistical agencies have developed a comprehensive range of classifications, and conduct regular reviews to achieve a balance between reflecting contemporary circumstances and consistency over time. The Australian Bureau of Statistics integrates or closely aligns its classifications with international standards such as the International Standard Industrial Classification of All Economic Activities (ISIC) and the Central Product Classification (CPC).

Other government agencies also produce standards for collections within their statistical domain. For example the Australian Institute of Health and Welfare publish the National Health Data Dictionary, which contains a set of definitions for use in Australian health data collections. The data dictionary is produced in consultation with a number of agencies including all Australian health departments, the Australian Bureau of Statistics, the National Centre for Classification in Health, the Department of Veterans' Affairs, the Australian Private Hospitals Association, representatives of the private insurance industry, Medicare Australia and the Australian Institute of Health and Welfare.

More information is available from the Statistical References section of this website.

## INTERNATIONAL STANDARDS AND CLASSIFICATIONS

There are many standards and classifications developed based on international recommendations. The United Nations Statistical Office, the International Labour Office, Eurostat, and other international and regional agencies produce international standard classifications.

### Appendix 4 - Confidentiality and Privacy

Legislation
Privacy Act
Privacy Commissioner
Freedom of Information Act
Ethics committees
Data Matching
Methods to Maintain Privacy
Techniques to Confidentialise Data

Agencies involved in the collection of data have a responsibility to ensure that procedures are implemented to ensure confidentiality and privacy. The primary goal is to ensure that there is no risk that an individual's or organisation's data is identifiable in output provided by a collection agency.

Confidentiality is ensured by making certain that information about individuals, households and businesses is not released outside the organisation and cannot be derived from released data.

## LEGISLATION

Legislative obligations governing the collection and release of information are outlined through Acts of Parliament and other government policy and guideline initiatives outlined below. These obligations seek to strike a balance between the need to collect and use demographic and sensitive information and the need to protect respondent/provider identity.

These obligations relate to the collection and dissemination of information in both public and private agencies.

## PRIVACY ACT 1988

The Commonwealth *Privacy Act 1988* enacted the principles outlined in the *Organisation of Economic Cooperation and Development Guidelines on the Protection of Privacy and Transborder Flows of Personal Data, 1984*. The 'Privacy Act 1988' provides protection for personal information that is handled by Federal and ACT government agencies.

Section 14 of the Privacy Act sets out eleven Information Privacy Principles that govern the conduct of Commonwealth agencies in their collection, management and use of data containing personal information. The Information Privacy Principles do not permit agencies to use or disclose, in identifiable form, records of personal information for research and statistical purposes, unless specifically authorised or required by another law, or the individual has consented to the use or disclosure.

The *Privacy Amendment (Private sector) Act 2000*, amended the Privacy Act 1988 to ensure that most private organisations are bound by the ten National Privacy Principles outlined in schedule 3 of the Act. The National Privacy Principles in the Privacy Act 1988 set out how private sector organisations and health service providers should collect, use, keep secure and disclose personal information. The principles give individuals a right to know what information an organisations holds about them and a right to correct that information if it is wrong.

The principles and guidelines of the Privacy Act are enacted either through state legislation or practice guidelines in each of the state and territories not covered by the Privacy Act, 1988.

For a copy of the Privacy Act, Information Privacy Principles, National Privacy Principles, data-matching guidelines or other related legislation see the Office of the Federal Privacy Commissioner web site.

## PRIVACY COMMISSIONER

The Office of the Federal Privacy Commissioner has responsibilities under the federal *Privacy Act 1988*.

The Privacy Commissioner gives advice:

- to individuals on their rights under the Privacy Act and related legislation
- about the Privacy Act and privacy issues more generally and promoting best practice in privacy standards
- to Federal and all state government agencies and other organisations on how to comply with the Privacy Act and related legislation.

The Privacy Commissioner also provides policy advice on privacy issues in response to written requests from Ministers, Federal and all state government agencies and the private sector, examines proposed legislation for privacy implications and conducts research into technological and social developments that affect individual privacy. The Privacy Commissioner investigates complaints from individuals about instances of interference with privacy, conducts audits of the personal information handling practices of Federal and all state government agencies and organisations under the Privacy Act and monitors the conduct of Federal government data-matching programs.

States enacted legislation and policies may be found at:

NSW - Office of the New South Wales Privacy Commissioner
NSW Health - Information Privacy: Code of Practice (December 1998)
Victoria - Office of the Victorian Privacy Commissioner
Western Australia - FOI
Tasmania - Information Privacy Principles
Queensland - Information Privacy Guidelines
Queensland Health Privacy Plan
South Australian Government Privacy Administrative Instruction
NT - Information Act (The Act) 2002

## FREEDOM OF INFORMATION ACT

The Freedom of Information Act 1982 provides individuals with the right to obtain access to documents (rather than information) in the possession of Ministers, departments and public authorities, other than exempt documents. This right is not restricted simply to documents in relation to personal information but allows an individual to request the amendment of records containing personal information that is incomplete, incorrect, out of date or misleading.

The 'Freedom of Information Act' lists several grounds upon which an agency may deny access: for example, where disclosure of the information may be injurious to law enforcement, or could threaten the safety of individuals.

Australian States and Territories have enacted their own Freedom of Information legislation. These may be found at:

NSW - Freedom of Information Act 1989
Tasmania - Freedom of Information - A User's Guide
Queensland - Freedom of Information Act 1992
Victoria - Freedom of Information Act 1982
Western Australia - Freedom of Information Act 1992
South Australia - Freedom of Information Act 1991
ACT - Freedom of Information Act 1989
NT - Information Act (The Act) 2002

## ETHICS COMMITTEES

In sensitive areas, such as health and medical issues, research ethics committees may exist or can be created directly for a specific purpose or project. The Guidelines Under Section 95 of the Privacy Act 1988, issued by the National Health and Medical Research Council provide a framework for the conduct of medical research using information held by Commonwealth agencies where identified information needs to be used without consent. In these situations, a Commonwealth agency may collect or disclose, in identifiable form, records for medical research purposes without infringing the 'Privacy Act' if the proposed medical research has been approved by a properly constituted Human Research Ethics Committee in accordance with the 'Guidelines Under Section 95 of the Privacy Act 1988'.

## DATA MATCHING

In its data matching guidelines, the Privacy Commissioner has defined data matching as "the large scale comparison of records or files of personal information, collected or held for different purposes, with a view to identifying matters of interest."

The Commissioner has issued advisory 'Guidelines for the use of data-matching in Commonwealth administration' for voluntary adoption by agencies conducting matching other than the programs specifically regulated by the 1990 Act. These guidelines therefore apply when the tax file number is not used in the matching process.

Data matching involves the bringing together of two or more data sets, at the unit record level, to form a composite record. It includes statistical matching as well as data linking (using identifiers) of data sets. The use of data matching techniques for statistical purposes is likely to increase. They have clear advantages in reducing provider load (compared with asking additional questions) and supporting longitudinal analysis. There are privacy issues, real and perceived, which need to be managed.

Statistical matching involves selecting core items that have been collected in different surveys or administrative data sets and using statistical matching techniques to synthesise records so that a richer data set can be used. These core items may be from different household surveys (e.g. age, sex, ethnicity, income, geographic location) or economic collections (e.g. industry, geography and employment size). Statistical matching techniques can be used to provide integrated micro-datasets for supporting micro-simulation and other analytical techniques.

Data linking (exact matching) involves drawing together datasets at the unit record level on the basis of a common identifier, for example Australian Business Number.

While linked datasets that bring together information from different agencies, potentially on a longitudinal basis, offer rich sources of statistical information, the feasibility of these datasets will depend on resolving related privacy, confidentiality and associated legislative concerns.

When record linkage of longitudinal data is to be performed, the record identifier should be used to ensure that once records are selected, they are followed over time. On the surface, it seems relatively straightforward to link units over time by matching the record identifiers at different points in time. However, potential complexities include tracking entries and exits, and identifying statistical units and appropriate record identifier links for complex units. If dealing with businesses, not all entries and exits will be legitimate births or deaths. Many will be the result of restructures, mergers, etc. Entries and exits need to be categorised so that appropriate units can be linked from one year to the next. Since businesses are continually changing the way they are organised/structured, it is necessary to ensure that like businesses are being linked.

## METHODS TO MAINTAIN PRIVACY

There are a number of ways in which the privacy of respondents can be maintained:

- code numbers, instead of names, can be used on the questionnaires to minimise the links that can be made between questionnaires and respondents

- suppression of personal identification information such as names, addresses and telephone numbers should be undertaken at the data processing or analysis stage

- questionnaires and data sources should be destroyed as early as possible in the process

- when creating tables, broad cross-classifications can be used to avoid cells with only a small number of contributing units.

## TECHNIQUES TO CONFIDENTIALISE DATA

The aim of making data confidential is to ensure that the confidentiality protection conditions are met while maximising the usefulness of the data outputs for analyses. Confidentialising can have a negative impact on the usefulness of data as some of the detailed data may need to be suppressed or modified.

## Tabular Data

Threshold rules and cell concentration rules can be established for tables. A threshold rule specifies the minimum number of units that must contribute to the value of a cell. Where the number of units contributing to the value of a cell is less than a pre-specified threshold value, the cell would be suppressed in order to prevent disclosure.

The cell concentration rule (also called a cell dominance rule) prevents the publication of cells where a small number of respondents contribute a large percentage to the cell total. For example, it may be decided that if any respondent/contributor accounts for a large percentage of a cell total, the cell will not be published.

There are several techniques that have been developed to minimise the risk of disclosure of information that can be traced back to the responding units. These techniques fall into three main categories, listed below:

- **Data Suppression**

  This technique simply involves not releasing information which may identify individuals in a cell. There are some simple automated suppression algorithms available for two dimensional tables. However, there are complex problems associated with tabulations of higher dimensions. To produce an efficient and practical automated system requires high resource input and much fine tuning. The primary suppression of sensitive or small cells may need to be complemented by the secondary suppression of other cells. For example, generally at least one other cell in a row or column containing a suppressed value also needs to be suppressed to prevent the original cell value from being rederived as the difference between the row or column total and the sum of the rest of the row or column values.

- **Data Rounding**

  Random rounding involves the technique of replacing small values that would appear in a table with other small random numbers. Since random rounding results in data distortion, it is not additive (additivity means that the table total, either between or within tables, are equal to the sum of the relevant cell values or subtotals). This technique can be unbiased if done in an appropriate manner. A value is biased if the expected value of the data after a confidentiality technique has been applied does not equal the value of the original entry it is replacing.

  Controlled rounding is a combination of conventional rounding and random rounding. Controlled rounding may result in additivity, unbiasedness, and reduction in data distortion (when compared to other rounding methods). However, this method may not provide consistency among tables.

- **Category Collapsing**

  Data items may be collapsed across classifications. Classifications which are very detailed, such as geography, country of birth, industry or occupation, can be collapsed down to a broader level.

## Unit Record Data (Micro Data) and Administrative Data

The generation and release of statistical information from administrative record collections usually includes data which are available at a detailed level both in terms of the characteristics of individuals and their geographic location such as postcode. Although personal information such as name and address may be removed, identification of individuals may occur by putting together information already known with the data provided.

The issue for agencies to address is what level of aggregation of data is required to avoid compromising the confidentiality of the individual's information and still produce meaningful data.

Different agencies have adopted different approaches to this issue. Individual agencies may have specific legislation which applies to their own particular situation. Where considered appropriate, legal advice should be sought prior to programming release of information to ensure the particular obligations of the agency are being met.

Two general methods can be used to ensure that unique identification cannot take place from micro data and administrative data:

### Data Reduction Methods

1. The identity of persons whom the information concerns is deleted together with any other information that identifies or is likely to identify any person.
2. Sampling. To reduce the probability of identification for some databases a sample of records rather than all records can be provided. This will introduce a further element of uncertainty as a user cannot prove that a unique record in a sample is also unique in the whole population without having some further information.
3. Ensuring that the populations for certain identifiable groups are sufficiently large.
4. Providing ranges instead of actual values for certain categories.
5. Top and bottom coding. Data for items such as number of persons under 15 in the household that can have unusual extreme values may be 'top coded' whereby all values beyond a certain level are placed in an open ended range.
6. Removing some of the variables from some respondents. Data items that were collected in the survey may be excluded from certain groups of records or from the file altogether.
7. Any cross-tabulations containing a count below a predetermined minimum are deleted from the data file to ensure that the identity of any person or organisation cannot be ascertained.
8. Removing the respondents from the file. A small number of records for which the above techniques have not been sufficient may be dropped from the file altogether.

**Data Modification Methods**

1. Adding 'random noise' to the data. For example, data for financial items such as income may be randomly adjusted up or down in a way that maintains mean values, or figures appearing in a table below a predetermined threshold can be randomly assigned an alternative value to disguise the real value.
2. Data swapping. For a small number of records sample weights and values such as year of arrival, industry or state of residence may be changed in a way that hides their uniqueness. The value in a problem record can be swapped with a value in another record (which may not be a problem record in itself) to limit the impact on aggregates.
3. Replacing the values for small groups with mean or median values.
4. Deleting information from some respondents and replacing it with imputed values. Imputation involves changing some of the responses or missing values on a record to ensure that a plausible, internally coherent record is created. This can be done automatically through matching person records containing missing values with donor person records based on simple demographic characteristics such as age and relationship in the family.

**Appendix 5 - Information Development Plans**

Information Development Plans (IDPs) aim to obtain agreement between key stakeholders about areas where further development is needed in a field of statistics. Long term, they hope to improve the quality, coverage and use of statistics within a particular field of statistics. An IDP is an agreement, developed as a collaborative effort between key stakeholders, that define the suite of information required to support policy in a particular field of statistics.

In simplistic terms, each IDP embodies three kinds of knowledge and shared commitment to statistical development activity:

- demand for information - a picture of the statistics that would, ideally, support informed design and evaluation of policy, other decision-making, research and community discussion.

- supply of information (and of raw data that might be used to create statistics) - a picture of the existing data pool that might satisfy the demand for information.

- agreed statistical development activity, identified through the comparison of demand and supply, which defines and priorities.
- information gaps (such as key variables arising in policy, decision making research or debate that have not yet been given statistical expression).

- information overlaps (such as variables for which competing or inconsistent measures are available).
- other information deficiencies (such as missing disaggregations by region or industry or sub-population, differing definitions or counting rules, or only rough approximations to the desired socioeconomic concept).

For an example of a well developed IDP see the AIHW Publications web site - National Public Health Information Development Plan 1999. This IDP was jointly prepared by the Australian Institute of Health and Welfare and the National Public Health Information Working Group.

## Steps in developing an Information Development Plan

### 1. Define the field of statistics

In developing an IDP for a field of statistics it is important to prepare a clear framework of what the particular field in question encompasses. For example, the National Public Health IDP defines public health as 'the organised response by society to protect and promote health and to prevent illness, injury and disability'.

### 2. Understand the debate

This step is about identifying the key questions, issues, debates and policy needs through consultation with stakeholders. Stakeholders would be comprised of federal and state government departments, research groups, industry bodies, lobby groups, private businesses and other users in the community. Formal consultations using new and established networks will be the key source of determining business need although opportunities will also exist for contributions through informal channels.

### 3. Define the desired information set

In this step an assessment is made on what information is required to satisfy the key questions, issues, debates and policy needs raised in Step 2. It should build a picture of the statistics that would, ideally, support informed design and evaluation of policy, other decision-making, research and community discussion. The key outcome is a list of variables which represents all of the information which could possibly be required and collected in an ideal world.

### 4. Assess the data pool

This step aims to identify the existence, structure, and quality of the available data pool for a particular field of statistics, and its potential uses. An assessment is required to determine which data sources within the pool can satisfy the information set defined in the previous step.

### 5. Decide what information would add value

The key aim of this step is to identify gaps, overlaps or deficiencies in existing data. The identification and agreement process should take place between the major stakeholders, which were identified in Step 2.

**6. Negotiate to create the information**

This is the stage at which the actual IDP is put together. This should be a collaborative process to create a plan or strategy for developing information for the field of statistics, based on the intelligence gathered and consultation processes in place. Decisions will need to made and agreed to in relation to which agencies will create the information.

**7. Evaluate**

An IDP is a dynamic process, as information needs change over time as do policy focus and issues. Stakeholders will need to develop a plan to regularly review the effectiveness of the IDP for their information needs.

### Appendix 6 - Statistical Skills

Staff Skills
Training

Agencies should have strategies in place to provide the skills necessary for professional and managerial competence in effective statistical operations and data management of the organisation.

## STAFF SKILLS

The conduct of statistical collections requires specific professional skills and managerial competencies. The main skills required are listed below. These skills may be contracted out to consultants. However, internal agency staff require basic skills in these areas at a level of competency which enables them to monitor and supervise external contractors.

- knowledge of project management principles and procedures and an ability to apply them to achieve well managed and cost effective statistical collections;

- awareness of data holdings and capacity to readily access all data holdings;

- understanding of the context within which a data collection is based, in terms of the broad frameworks for the statistics;

- understanding of and ability to apply statistical frameworks, classifications, standards and concepts;

- understanding of sample frame creation and maintenance procedures relevant to a collection and the impact these have on the statistics produced;

- capacity to identify and apply the most appropriate data collection vehicles, including the use of administrative sources;

- ability to develop appropriate data collection instruments. This includes an understanding of, and ability to apply, form-design principles and an ability to apply appropriate testing techniques to evaluate the quality of reported data.

- understanding of sampling procedures relevant to a collection, the impact these have on the statistics produced and implications for collection and processing operations.

- understanding of quality issues relevant to the use of data from administrative systems;

- ability to communicate effectively and be sensitive to the burden on providers, and to understand how technology can support effective provider management;

- expertise in developing, within the context of corporate standards, appropriate and efficient provider management, data capture and processing mechanisms;

- ability to apply good practice in the effective use of relevant corporate data storage, processing and dissemination facilities;

- ability to develop effective dissemination strategies, including the determination of content of disseminated material, based on sound analysis of the data; and knowledge of the interactions between data structures and technology to get the most efficient dissemination processes;

- expertise in identifying and commenting on the most significant features of the data, and the reasons for unusual results. To do this requires the ability to look at data sets from both time series and cross-classified perspectives and to evaluate data sets at varying levels of detail. It also involves knowledge of how methodological aspects (e.g. sampling, seasonal adjustment, price deflation, etc) affect data sets;

- ability to service client requirements for information efficiently and effectively; and

- for continuing collections, ability to evaluate and improve current systems, processes and procedures to increase efficiency and/or quality of data produced, i.e. the ability to identify and implement continuous improvement activities.

## TRAINING

Skilled staff are essential to achieve and maintain a high quality data collection. Agencies should undertake a regular skills stocktake of staff to identify shortages in statistical skills so that appropriate measures may be implemented. This may be through training courses, the recruitment of appropriately skilled staff, contracting of external consultants or obtaining expertise from elsewhere in the organisation.

Training methods can include formal statistical training courses which can be supplemented with on the job guidance and supervision. Detailed documentation on procedures, classifications, coding and systems should be provided to new staff and can also be used as a reference for existing staff. More information on available training courses can be obtained from NSS Statistical Training Services.

**Appendix 7 - Quality Declaration and Assesment**

Relevance
Accuracy
Timeliness
Accessibility
Interpretability
Coherence

Data accuracy and reliability can be assessed by using quality assurance measures. The Australian Bureau of Statistics uses a template to describe the quality of a data source, where different issues are addressed under the different dimensions of the data quality framework. The completed template is called a Quality Declaration.

Each characteristic can be assessed using the following criteria:

- The data collection significantly falls short of requirements
- The data collection is sufficient with some areas of reservations
- The data collection is sufficient for the requirements
- The data collection significantly exceeds requirements
- There is insufficient information to judge the suitability of this characteristic.

Listed below is a description of each dimension of data quality as well as the template which provides a number of relevant data characteristics which need to be assessed against the criteria listed above. For example, for 'relevance', some of the characteristics are the coverage, reporting unit, frame, classifications and concepts. A definition for each characteristic is provided in the third column. In the last column, there is an explanation of how that characteristic is important from a user assessment perspective. For example, if the 'coverage' of the collection excludes people or groups that the user is interested in, the user will need to make some judgements about how these exclusions impact on their decision making capability.

The collection area can also use the template and the above assessment question to assess whether the collection needs modifying or not. For example, if the assessment indicates that the collection significantly falls short of user requirements, then they be able to modify certain aspects of the collection, such as the coverage, sample size etc. Conversely, if the assessment indicates that the collection is exceeding requirements, then maybe savings could be made by reducing the sample size or coverage of the collection.

## RELEVANCE

The **relevance** of statistical information reflects the degree to which it meets the real needs of clients. This is addressed in the Quality Assessment by:

- looking for mismatches in coverage, classifications, concepts and data items between what the data collection provides and what the user requires; and
- understanding who the respondents are and how the information is collected.

Looking for mismatches is important because a mismatch tells us that the data collection is not measuring exactly what user wants to measure. As such, it is important to understand the potential impact of the mismatch on the decisions that the user wishes to make.

Understanding who the respondents are (e.g. universities) and how the information is collected (e.g. electronically via e-mail or the Web) is important because this assists in better understanding the limitations of the resulting data.

| **Data Characteristic** | **Questions to be answered by Collection Manager** | **Definition/Explanation** | **Relationship to Quality Assessment** |
|---|---|---|---|
| Coverage | What is the population actually covered by the data collection? | Coverage includes both the geography covered by the data collection (e.g Victoria) and any other rules used to identify whether a unit is included or not (e.g. exclude people 15 years or under, exclude non-residents) . | If the population covered by the data collection is different to the population you are interested in, then the following questions need to be asked:<br>□ For parts of the population that you are interested in, but are not available in the data collection, are they likely to exhibit different characteristics?<br>□ Can you subset out parts of the population that you are not interested in, but are included in the data collection? If not, are these additional units |

| | | | |
|---|---|---|---|
| | | | likely to exhibit different characteristics?<br>In answering these questions, it is also important to remember the impact on totals as well as averages. For example, missing out on a part of the population is likely to mean that the totals will be too low. |
| Reporting Unit | Who collates and provides the data? | The reporting unit describes who actually provides the data. In some cases, the reporting unit will also be the unit of interest. However, this will not always be the case (e.g. universities might report course information on students).<br>In those cases where the data are provided and collated by different people, please details of both. | The reporting unit is important as the information collected will generally be from the perspective of the reporting unit. For example, collecting information on fields of study from students and the institution they are studying at may well produce different results. |
| Frame | How is the list of potential respondents compiled?<br>Are there any data quality issues associated with the frame (e.g. new units, defunct units, duplicates, age of frame)? | The results from a data collection are highly dependent on the list used to identify who should respond to the data collection. The quality of this list will have a strong impact on the quality of the data. | For example, a list prepared using the White Pages would exclude households without home telephones or with silent numbers. The issues here are similar to those identified for coverage. |
| Classifications used | What are the key classifications used? | A classification is set of defined groupings or categories - based on common relationships - into which all members of statistical units can be divided or arranged. These groupings or categories can be ordered systematically, are mutually exclusive and exhaustive, and are based on one or more data items. | If the classifications used in the data collection do not match up with requirements, then it is important to consider the potential impact of this. For example, a difference in industry classifications may mean that you are unable to exactly measure the industries you are interested in. As with coverage, it is important |

|  |  |  |  |
|---|---|---|---|
|  |  | Examples of classifications include: State, Industry; Highest Level of Educational Attainment; Age (in 5 year groupings); and Country.<br>In those instances where classifications used correspond to industry, national or international standards, this should be indicated. | then to assess the likely impact of this mismatch. |
| Concepts used | Describe any key concepts addressed in the data collection. | A concept in the context of a data collection usually refers to an issue which is often difficult to measure directly (e.g. well-being, some economic concepts) or needs to be derived through several data items (e.g. unemployment, disability).<br>Often the key concepts are the key issues which the primary user is seeking to measure in the data collection. | If the data collection is not measuring the exact concept you are interested in, it will be necessary to assume that the concept you are interested in would produce similar results to those in the data collection, had it been measured. The greater the difference in the concepts, the more tenuous this assumption becomes and the greater the danger that decisions will be made using data which are not conceptually relevant to your needs. |
| Key data items | What are the key data items collected? | A data item is a particular characteristic which is measured or observed. There are two main types of data items:<br>□ Parametric data items are quantitative measures and have both an associated unit of quantity (e.g. $, hectares, hours) and an associated type (e.g. flow, stock, index, movement).<br>□ Classificatory data items are described in terms of a category (e.g. industry, state, country of | For the collection to be useful, it needs to collect the information you are interested in. Mismatches in data items will lead to similar problems as mismatches in concepts or classifications. |

| | | | |
|---|---|---|---|
| | | birth) rather than using a quantitative or numerical measure. | |
| Mode of data collection | What mode of data collection is used? | The mode of the data collection describes the method used to collect data. Examples include:<br>□ e-mail;<br>□ web;<br>□ Computer Assisted Telephone or Personal Interview; and<br>□ Personal Interview. | The way the data are collected may lead to certain limitations in the data, often relating to the coverage of the data collection or the type of information that can be collected using that mode (e.g. personal interviews may cause problems with sensitive questions, but allow the interviewer to better clarify issues with the respondent). |
| Intended audience and purpose | For what purpose(s) is the data collection run?<br>Who is the primary intended audience for the data? | The intended audience are the primary users of the data collection. In most cases, the data collection will have been designed specifically to meet the needs of these users.<br>The purpose of the data collection is defined as the primary use for which the data will be used by the intended audience. | While a Quality Assessment is not required for this characteristic, this assists in providing the user with an understanding of the broader context of the data collection. |
| Owner of data collection | Who is responsible for managing the data collection?<br>Who is responsible for deciding on the data items collected? | The data collection manager is defined as the person or position responsible for the operation coordination and running of the operational aspects of the data collection.<br>In addition, a person or group of people will be responsible for deciding which data items are collected or included. This may differ from the data collection manager. | Once again, a Quality Assessment is not required for this characteristic. |
| Authority | Under what (and whose) authority and/or legislation is the data | This provides information relating to expected response rates and the | Not assessed in Quality Assessment. |

| | | | |
|---|---|---|---|
| | collection run? | general context under which the respondent is required to provide the data. For example, the quality of information collected under an Act of Parliament for the provision of federal funds might be expected to differ from that collected from university administrative records provided on a purely volunteer basis. | |

## ACCURACY

The **accuracy** of statistical information is the degree to which the information correctly describes the phenomena it was designed to measure. As such, it is important to consider issues of both sampling error and non-sampling error (where applicable).

Issues such as mismatches in coverage or classifications may also be considered here, but they are addressed primarily under Relevance.

For the Quality Assessment, the user needs to consider whether the accuracy of the data collection will be sufficient to meet their needs. If not, they then need to consider the impact of using the data. This may mean that decisions will be made using data from the data collection, when the underlying information that they are interested in could be significantly different. In other words, the data may be misleading, resulting in poor decisions.

| **Data Characteristic** | **Questions to be answered by Collection Manager** | **Definition/Explanation** | **Relationship to Quality Assessment** |
|---|---|---|---|
| Level of sampling error (applicable to survey samples only) | What are the relative standard errors for key data items? Include relative standard errors of key data items also for key subpopulations. | Sampling error reflects uncertainty in the true population value because information was collected from only a sample of the population. This is often measured as the relative standard error (i.e. standard error of the estimate as a percentage of the estimate). This can be used to identify a range of values that the true value is expected to lie between (e.g. 95% confidence interval). | If the range of values is high, this can impact on the decisions based on the data. For example, if you knew that the unemployment rate was in the range between 0% and 20%, would this restrict the type of policy decisions that you would be comfortable making? |
| Response | What is the response rate? | The response rate is | In most instances, an |

| | | | |
|---|---|---|---|
| rates | What steps are in place to attempt to maximise the response rate? | calculated by dividing the number of responding units by the number of units which were selected and were in scope of the data collection.<br>Examples of methods used to maximise the response rate include (but are not restricted to):<br>□ use of primary approach letters;<br>□ interviewers well trained in establishing a rapport with respondents or the design of respondent-friendly questionnaires;<br>□ informing respondents how the results of the data collection will benefit them; and<br>□ detailed call back strategies. | assumption is made that the non-respondents would have provided similar information to the respondents. However, the non-respondents may in fact be quite different to the respondents, so the data will be biased to reflect those units which have responded. For example, imagine a data collection on university students where all the overseas students failed to respond. Had the overseas students responded, different conclusions may have been reached.<br>In interpreting the response rate, it is important to consider how your conclusions based on the data may have changed if the non-respondents had responded very differently to the respondents. This is often best handled using a sensitivity analysis approach.<br>In completing the Quality Assessment, first consider how much the data would be likely to change and then consider how that might impact on any resulting conclusions or decisions made.<br>Understanding the steps for maximising the response rate should provide some insight into the potential for non-response bias. |

| | | | |
|---|---|---|---|
| Adjustments to data | What methods are in place for edits and data validation?<br>What data items have more than 10% of units with missing values or have been edited or imputed?<br>For imputed data items, approximately what percentage of units have been altered on the basis of editing or imputation?<br>Are the data subject to large revisions? | Editing is the process of checking data records to ensure that they contain valid entries and changing the records where they do not, whereas imputation is the process of estimating data for individual records which have not been completed. Data validation is a general term for methods used to check that the data appear correct. | The concerns with high levels of editing and imputation is similar to the concerns associated with high levels of non-response. That is, how much are our decisions being influenced by data which didn't come directly from the respondents but were estimated?<br>Similarly, if the data are subject to large revisions, there is a high degree of uncertainty about what the final data will actually be. Consider how much the data might change due to revisions and whether the revised data would lead to different decisions. |
| Other data issues | Are there any other issues that might impact significantly on the accuracy of the data? | Other issues may also impact on how well the data being collected actually measures what it is supposed to measure. Examples include:<br>□ different levels of data quality for different data items in administrative collections;<br>□ sensitive information; and<br>□ recall bias. | Other issues, such as those listed here, can also influence the data collection's ability to accurately measure what the user actually wants to measure. For example, the respondent may not be able to provide the information with any degree of certainty, as they cannot remember the details sufficiently or they are being asked to provide an opinion on something on which they feel they do not have sufficient information. This information also needs to be considered as part of the accuracy of the data. |
| Level of training | What is the level of training received by those involved in the collection design and operation (i.e. | Poor training can cause significant problems with the ultimate quality of the data. For example, | In making a Quality Assessment, the user needs to consider whether the level of |

| | | | |
|---|---|---|---|
| | questionnaire design, systems used to collect information, systems for editing and processing the data, etc.)? | questions could be misleading or ambiguous so the respondent may not have interpreted the questions as was originally intended. Similarly, poor training for data processing could lead to errors being introduced at data entry. | training is sufficient for the data collected. This will be related to the nature and complexity of both the data collection procedures and the data to be collected. |
| Comparability in data values with related data sources | How does the data collected compare with similar data sources? | Comparability in data values with other data sources offers some insight into whether the data seem to be measuring what the user is interested in (noting that the user's requirements may be sufficiently different to prevent the use of the other data sources). | For this characteristic, the Quality Assessment focuses on whether a possible lack of comparability between the data values from this data collection and other related sources is sufficient to cause some concern with the data collection. |

## TIMELINESS

The **timeliness** of statistical information refers to the delay between the reference point (or the end of the reference period) to which the information pertains, and the date on which the information becomes available.

| **Data Characteristic** | **Questions to be answered by Collection Manager** | **Definition/Explanation** | **Relationship to Quality Assessment** |
|---|---|---|---|
| Recency of data | What was the last reference period for the data collection?<br>How often is the data collection usually run?<br>When is the next data collection expected to occur?<br>When are data to be extracted from the administrative system? | The reference period refers to span of time to which the data refers. This may either refer to a single point in time or a span of time. | For this characteristic, the Quality Assessment is asking about the suitability of the timeliness of the data. If circumstances are likely to have changed significantly since the last time the data were collected (e.g. internet usage) and the data needs to reflect the current situation, there will be problems comparable to those experienced under the relevance and accuracy dimensions - the data may not be |

| | | | |
|---|---|---|---|
| | | | measuring what the user wants to measure which may lead to inappropriate decisions using that data.<br>Thus, it may be concluded that the value of the data is limited given that the data are no longer relevant to the current situation. |

## ACCESSIBILITY

The **accessibility** of statistical information refers to the ease with which it can be accessed by users. This is addressed in the Quality Assessment by considering:

- ease of accessing the data; and
- knowledge that the data exist.

This impacts on decisions regarding whether the data collection is an appropriate data source, with respect to ease of obtaining the data, its security and the impact on any dissemination of results.

| **Data Characteristic** | **Questions to be answered by Collection Manager** | **Definition/Explanation** | **Relationship to Quality Assessment** |
|---|---|---|---|
| Ease of getting data | What is the average time taken to fulfil a data request?<br>What data are readily available on the Web?<br>What publications are available, and where are those publications available?<br>What is the associated pricing policy? | A data request will generally refer to a request for tabulated data. The complexity of the data request may vary for different requests so consider the average time required to meet a request of 'average' complexity.<br>The pricing policy is the set of rules or guidelines for determining the cost for a user to purchase data. | Even having received permission to access, it might prove too difficult to get the data in a suitable form or it might take too long to get the data. Similarly, access to the data may prove to cost too much given your available resources. |
| Knowledge data exist | How are people internal to the department made aware about the existence of the data? | Knowledge that the data exist is an important aspect of the accessibility of a data collection. This | In the comments field, the user should indicate how they became aware of the data and how easy it was |

| | | | |
|---|---|---|---|
| | | includes how information on the data collection is made available both internally (e.g. on the Collection Management System with the ABS or externally (e.g. on the Web or hardcopy publications in most libraries). | for them to locate the data. It is expected that ABS data are listed and documented on the Collection Management System. |

## INTERPRETABILITY

The **interpretability** of statistical information reflects the availability of the supplementary information and metadata necessary to interpret and utilise it appropriately. Interpretability has been addressed in the Quality Assessment by asking the questions:

- Is there sufficient information to make an informed Quality Assessment on all characteristics?
- How easy is it to obtain more information about the data and data collection if required?

If there is insufficient information to understand properly how well the data meets the user's specific needs, then they are in danger of using inappropriate and/or misleading data to make important decisions.

| **Data Characteristic** | **Questions to be answered by Collection Manager** | **Definition/Explanation** | **Relationship to Quality Assessment** |
|---|---|---|---|
| Level of documentation for data collection | Has this Quality Declaration been 'signed off' by the area in charge of the data collection?<br>What more detailed information is available regarding the data collection? | The Quality Declaration is the document as described in this appendix. The level of documentation should be aimed at providing sufficient information for someone without previous knowledge of the data collection to complete a Quality Assessment (without using the assessment of 'insufficient information').<br>More detailed information might be available through other sources, such as user guides, a web site or other documentation | The Quality Assessment for this characteristic makes an assessment as to whether the level of documentation in the Quality Declaration is sufficient. Insufficient information to make a Quality Assessment means that there is uncertainty regarding the data quality for that characteristic. As such, any decisions using the data which are affected by that characteristic will be based on data on dubious quality and may lead to inappropriate decisions being made. |

| | | | |
|---|---|---|---|
| | | maintained by the data collection manager. | The Quality Assessment should indicate that the level of documentation is insufficient for those characteristics which have been rated as "there is insufficient information to judge the suitability of this characteristic". The comments field should indicate which characteristics have received this assessment. |
| Internal accessibility of documentation for data collection | Is the Quality Declaration readily available within the department? | The Quality Declaration should be available for all potential users of the data within the department to access, in case they need to review available data sources for a given need. Ideally the Quality Declaration should be stored on a corporately endorsed (standard) storage medium for documentation on data collections. | In the comments field, the user should indicate how easy it was for them to locate the Quality Declaration. |
| External accessibility of documentation for data collection | Is this Quality Declaration available to people outside the Department (Web or Other - please specify)?<br>Is more detailed information available on the Web (Web or Other - please specify)?<br>What level of documentation is provided in publications? | This characteristic refers to the availability of the Quality Declaration to people who do not work within the department. More specifically, these people include anyone who might be interested in understanding the quality of the respective data collection (e.g. academics, policy analysts in other departments).<br>Possible methods for external accessibility would include the inclusion of the Quality Declaration on a web site or in publications released to the general public. | For this characteristic, the Quality Assessment comments on whether documentation will be made sufficiently available for those outside the department.<br>For users within the department, this assessment draws on whether it is important that people outside the department are able to access the documentation (e.g. to support published data). This would also alleviate the degree to which the area managing the data collection needs to be called upon to |

| | | | |
|---|---|---|---|
| | | | answer questions about the data collection. |

## COHERENCE

The **coherence** of statistical information reflects the degree to which it can be successfully brought together with other statistical information within a broad analytic framework and over time. This is captured in the Quality Assessment by focusing on changes over time to the data collection as any such changes will impact on any interpretation of how things may have changed over that period. For example, a perceived change in results between two time periods might simply reflect a change in definition. Thus, it is important to know when these definitions have changed and how much they have changed, and considering the potential impact of those definitional changes on the data.

| **Data Characteristic** | **Questions to be answered by Collection Manager** | **Definition/Explanation** | **Relationship to Quality Assessment** |
|---|---|---|---|
| Consistency of classifications over time | List any changes in key classifications over time | To try and maintain or improve the general relevance of a classification, they are often reviewed and updated over time. Examples of classifications which are subject to review include:<br>□ Statistical Local Areas (geographic);<br>□ Collection Districts (geographic);<br>□ Industry; and<br>□ Countries. | The Quality Assessment should consider the changes in the key classifications in the specific context of the user's requirements. Some classifications may not be relevant to your needs or the changes may be minor compared to your needs. Alternatively, some changes may cause major problems in comparing data over time. |
| Consistency of concepts and methodology over time | List any changes in key concepts and methodology over time. | To try and maintain or improve the general relevance of a statistical concept, they are often reviewed and updated over time. For example, the concept of employment as measured may have changed over time. Similarly, other concepts such as innovation have evolved over time as more research is done in their respective fields. | The issues associated with this are the same as those listed above for assessing the consistency of classifications over time. |

| | | | |
|---|---|---|---|
| | | Similarly, changes in the collection methodology may impact on the resulting data. Examples might include changing the data collection methodology or the questionnaire. | |

**Appendix 8 - Using Quality Declarations to Manage Risks**

Risk Management Strategy
Stage 1 - Understanding the Risks
Stage 2 - Mitigating Risks
Stage 3 - Plan for the Unexpected

Project Management should provide for a clear plan tying together the statistical activity's objectives, scope, intended outcomes, outputs and the key issues of stakeholders. A project plan allows the manager of a statistical activity to coordinate the various stages of the statistical cycle, documenting the flow and timing of the tasks involved to ensure that all aspects have been considered, including dependencies, quality assurance, resourcing, project governance, and to determine at any stage whether the statistical activity is running to schedule and budget.

An effective Project Management Plan should cover the following issues:

- ensuring all tasks are included and sufficient time is allowed for them

- identifying tasks which are dependent upon a previous task(s) and who is responsible for each task

- identifying times, or points in the timetable, to assess progress of the whole activity and ensure the objectives are being met.

A Project Management Plan should incorporate a Risk Management Strategy. This is discussed in more detail below.

## RISK MANAGEMENT STRATEGY

Risk management is increasingly recognised as a central element of good corporate governance and as a management tool to assist strategic and operational planning. Risk management is a set of activities concerned with identifying and understanding potential risks affecting the organisation's or a program's major function, analysing their consequences (probability and impact), and devising and implementing responses (controls) so as to ensure that corporate and business objectives are achieved. Risk management is about identifying opportunities and avoiding or mitigating losses**.**

In regard to statistical activities, potential risks include:

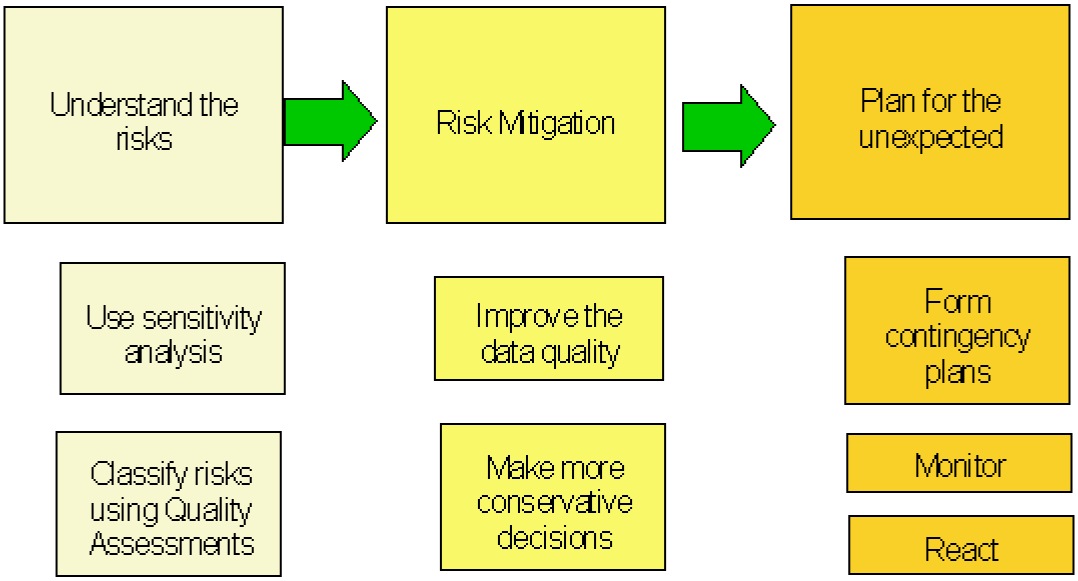
- applying methodologies which are considered questionable
- data collected doesn't match with the data need
- analysis is not viewed as objective
- charging policy for data prevents access by certain groups
- staff are not sufficiently trained to undertake the required statistical work
- quality control measures are not implemented
- scope is poorly defined
- computer data are lost
- committees and boards do not operate effectively.

### Example: Incorporating Risk Management into Decision-Making

Risk management can be incorporated into the decision-making process if decisions are being made based on data. If the data quality is poor, then the risks of making a poor decision using those data are greater. Conversely, if the data quality is high, then greater confidence can be placed in the information being used to make an informed decision.

To facilitate the decision-making process for data, risk management has been subdivided into three stages which are broadly summarised in the diagram below.

**Figure 1 - Risk Management**



## STAGE 1 - UNDERSTANDING THE RISKS

The first step in risk management is to understand the risks, and in using statistical data this involves two key steps: sensitivity analysis, and classifying risks using Quality Assessments.

The purpose of sensitivity analysis is to identify the various levels of risk associated with using a specific data source for a given data need. This is achieved by examining each of the data characteristics identified in the Quality Declaration (See Appendix 7 - Quality Declaration and Assessment) and trying to understand the potential impact on the underlying decision of using the data.

As a result of this analysis, each characteristic is classified according to the degree of match between the data need and the data source, ranging from 'the data collection significantly falls short of requirements' to 'the data collection significantly exceeds requirements'.

Sensitivity analysis can be as simple as going through a process of asking two key questions:

- How different would the data need to be for me to make a different decision?

- How likely is it that the data would be that different as a result of the mismatch between my data need and the data source?

## STAGE 2 - MITIGATING RISKS

This step involves investigating options for reducing the level of risk. This example has identified two such avenues:

1. Improving the data quality; and
2. Making more conservative decisions.

### 1. Improving the data quality

Improving the data quality primarily deals with looking for opportunities to improve the match between the data need and the data source. There are a range of opportunities to improve the quality of the data source.

- Apply a data collection to only part of the problem.

  This option accepts that the data source is sufficient for answering part of the question, but not all of it. For example, the scope of the data collection might only cover specific States but information is required for all States and Territories.

- Accessing multiple data sources.

  Accessing multiple data sources provides the opportunity to either validate existing data sources or use each data source for its respective strengths and alternative data sources to cover the weaknesses. This may be particularly useful when a data collection has been found to be suitable to apply to only part of the problem. Multiple data sources may also provide information for small area estimates where these are not otherwise available.

- Deciding more information on the data collection is required.

  Risks that are identified in the Quality Assessment with the rating of 'There is insufficient information to judge the suitability of this characteristic' can be mitigated by undertaking further research to provide sufficient information to assess the suitability of the characteristic in question. These risks can then be adequately assessed and identified.

- Modifying an existing data source

Any decision to modify the data collection will need to assess the cost of implementing changes against the benefits. The Quality Assessment should provide a key indication of the type and extent of modification required. The data characteristic measured against the Quality Assessment, as set out below, will identify the issues that need to be answered.

1. The data collection significantly falls short of requirements: the data collection may require a significant structural change to better meet data need requirements. For

example, the scope of the data collection might be changed to include a broader geographical scope, the sample size might be increased to meet specific user requirements, or the databases might be improved to enable easier access to the data.

2. The data collection is sufficient with some areas of reservations: the assessment has already identified that the data collection is already nearly sufficient to meet needs so this may indicate a smaller structural change to the collection or minor alterations to collection processing or procedure.
3. The data collection is sufficient for the requirements: while the data collection may meet the specific needs of the user, the user may have identified some potential areas for improvement, which may involve only marginal additional costs or allow for the data to be used more effectively across a wider range of purposes.
4. The data collection significantly exceeds requirements: this rating may indicate that some savings might be achieved by reducing what is being offered. For example, the amount of editing might be reduced or the sample size might be reduced.
5. There is insufficient information to judge the suitability of this characteristic: this reflects a need to improve the level of documentation available to provide a better mechanism to assess the quality of the collection and provide a suitable assessment.

- Deciding a new data source is required

It may be decided that the risks associated with existing data sources are too high and cannot be sufficiently mitigated. In this case, it might be necessary to develop a new data source. The same process of assessing the proposed data source against the specified data need using the data quality framework should be followed.

While the above options are available to enhance the overall quality of any given data source, resource limitations may mean that compromises need to be made to achieve an 'affordable level of quality'. However, in making these compromises, two issues need to be considered:

- Where should compromises be made?
- Once compromises have been made, will the data still meet data requirements sufficiently?

## 2. Making more conservative decisions

Having identified the risks, it is important that the underlying decision takes these risks into account. In other words, these decisions should take into account the quality of the data as well as the values of the data.

It is difficult to provide specific options here, as they are dependent on the underlying decision and the corresponding areas of risk. For example, a decision may need to be made on whether an allocated budget will support the introduction of a localised program to support unemployed persons. Rather than initiating the complete program, more conservative decisions might include implementing localised trials, restricting eligibility criteria for program assistance (e.g. based on age or length of unemployment), delaying a decision pending more information or even deciding to use the money to expand existing programs.

## STAGE 3 - PLAN FOR THE UNEXPECTED

### 1. Form Contingency Plans

In Understanding the Risks, Quality Assessment was used to classify the risks. The areas which were identified as higher risks are the same areas where contingency plans are required (unless the risks were later mitigated sufficiently).

Contingency plans are simply strategies of what to do if certain risks are realised. For example, a low response rate for a survey generates a risk that the survey results are significantly influenced by non-response bias. As a result, inappropriate decisions might be made on the basis of the biased results. A contingency plan will have answers in place to solve problems associated with a higher than expected degree of non-response bias.

These contingency plans should relate to the underlying decision. Using the unemployment program example referenced earlier, a contingency plan might be to reduce or drop the program if uptake proves to be much lower than expected. Similarly, the program may be moved to a different location or eligibility criteria expanded. These plans can be very similar to those considered early at the risk mitigation stage. However, instead of mitigating the risk immediately through making a more conservative decision, the decision might not fully take into account the associated risks. Rather, the risk mitigation option would only be implemented if further information suggested that the risks had been realised.

### 2. Monitor and React

Monitoring is a key part of planning for the unexpected. Having formed contingency plans, it is important that the information is available which will trigger these contingency plans into action.

While it may be possible to continue to monitor data from a regular survey or an ongoing administrative data system, this will not always be possible. As such, other ways to monitor the impact on the underlying decision should also be considered. For example, monitoring budgets would assist in avoiding overspending budget allocations. Similarly, a decision to run specialised training programs for the unemployed would benefit from monitoring both participation levels in training programs, participant comments on the training and overall levels of unemployment.

## Glossary

### A

Acceptance sampling
The use of statistical sampling for determining the acceptability of work lots or groups of items. Based on inspection of a sample or samples, acceptance or rejection of the work or items is determined according to prescribed decision rules. (See also: Quality control).

Accessibility
A quality measure relating to the ease with which statistical data and published estimates can be retrieved, used and understood.

Accuracy
A quality measure relating to the degree to which the statistical information correctly describes that which it was designed to measure.

Administrative data
Information primarily collected for the purpose of record-keeping, which is subsequently used to produce statistics.

Administrative system
The store in which records are held.

Agency
A discrete organisation which is controlled by a Federal or State/Territory Government, including Departments, Offices and Statutory Authorities.

Aggregation
The grouping of units into categories and the summing of values within these categories.

Allocation
The manner in which the total sample is distributed to different parts of the population by the sampling plan. This determines the number of units to be selected from each part.

Analysis
The summary and interpretation of data in a collection for the purpose of measurement against defined objectives.

Analysis unit
The entity about which data are collected, for example person, household, business, et cetera. See Unit.

ANAO
Australian National Audit Office.

Anonymity
In a survey, anonymity exists if the identity of each respondent who has returned a completed questionnaire is not known to anyone other than the respondent and no respondent can be identified by inference from the results of the survey. If code numbers or other identifiers are put on the questionnaire when sent out, anonymity does not exist. (See also: confidentiality).

ANZSIC
Australian and New Zealand Standard Industrial Classification: a system for identifying and grouping all producing units in Australia on the basis of industries (for example manufacturing, mining, education) to permit compatibility of data.

ASCO
Australian Standard Classification of Occupations: a system for identifying and grouping all occupations in Australia on the basis of the tasks, duties and responsibilities associated with the occupation.

Attitude scale
(See: Rating scale).

Attribute
A characteristic of a person, object or concept which can be described only in terms of categories (e.g., marital status, gender, hair colour, etc.), rather than being described in quantitative or numerical units. (See also: Variable).

Auxiliary information
Related information from a number of sources, used in the design of a particular sample

### B

Benchmark
A point of reference from which subsequent measurements or observations of the same items of interest may be made.

Bias
The tendency, during any step in a survey, to systematically favour or

give advantage to answers or findings which will cause resulting estimates to deviate from the true value.

## C

Case study A method of teaching or study in which the relevant experiences of a person, organization, etc., are recorded and analyzed, in order to gain insights into a type of conduct which is assumed to be common to a whole group or category of people, organizations, etc., or which illustrate a particular type of situation.

Census The complete counting of all individuals in a target population, for example, a manufacturing census collects information from every known manufacturer. When capitalised, "Census" usually refers to the national Census of Population and Housing.

Classification A structured set of mutually exclusive defined groupings or categories, based on common relationships, into which all members of a defined population can be systematically arranged or divided.

Classificatory item A data item which is described in terms of a category (for example industry, state, country of birth) rather than by a quantitative or numerical measure. See Parametric item.

Clean The state of a data item, survey response or dataset that has been fully edited and is ready for analysis or dissemination.

Cleansing A process of improving the quality of a dataset by checking for consistency and invalid codes and resolving reporting anomalies.

Clerical editing The manual checking of responses in a survey before data entry.

Cluster analysis A method for identifying data items that closely resemble one another and assembling them into groups or clusters.

Coding The assignment of numbers, letters or symbols to data items to facilitate processing and analysis.

Coefficient of variation The standard error of an estimate, expressed as a ratio or percentage of the estimate (e.g., the standard deviation of a distribution divided by its mean). It is a measure of the relative dispersion of distributions and is useful because it is independent of the unit of measurement by which the basic variable is measured.

Cognitive test A test designed to understand respondent thought processes and identify errors that may be introduced while completing a survey questionnaire. The information obtained is used to make improvements to the questionnaire.

Coherence A quality measure relating to the degree to which statistical information can be brought together with other statistical information within a broad framework and over time.

Cohort study A study at two or more points in time of certain characteristics of the member of a sub-population which have some particular common attribute, e.g., a study of persons born in a specific year, a study of the graduating class of a specific year, etc. (See also:Longitudinal study).

Collection See Data collection, Statistical collection.

Collection instrument The medium or vehicle by which information is obtained from respondents, for example paper form, electronic form, interview schedule, computer assisted interview.

Collection methodology The framework of principles and procedures used in gathering data on a particular topic. See Methodology.

Collection unit See Reporting unit.

Comparability The ability of statistics within the one survey, over time, or across collections, to be compared with one another.

Computer assisted telephone interviewing (CATI)
A type of telephone interviewing in which the interviwer keys in answers to questions as they are received, onto a data-entry keyboard. A viewing screen automatically displays the appropriate question to be asked next or, if warranted, an error message.

Confidence interval
A measure of the probability that the true value for a population lies within a given range of values.

Confidentialise
To present statistical data in such a way that information about individual respondents cannot be derived from macrodata.

Confidentiality
The protection of information provided by respondents, and assurance that information about individual respondents cannot be derived from macrodata.

Consistency
The extent to which data items, when considered together, make sense and are not contradictory.

Consistency edit
A check that a precise arithmetic relationship between two or more data items is obeyed, for example that a reported total equals the sum of the reported components.

Correlation
A numerical or algebraic relationship between two variables.

Coverage
A term used in sampling with several related meanings. (1) The individual members of a target population about which information can be obtained. (2) The part of a population covered by a sample (for example 50% coverage indicates that 50% of the target population was examined). (3) The extent of the material collected from sample members.

Cross-sectional survey
A survey which collects information about characteristics of and relationships between units at a single point in time.

Cross-tabulation
The categorisation of data by two or more variables, usually presented in tabular format.

Cross classificatory variable
A variable which can be grouped by two or more other variables, for example, person by age and sex.

## D

Data
A representation of facts, concepts or instructions in a formalised manner, suitable for communication, interpretation or processing.

Data capture
The initial recording of data, for example responses to a questionnaire on paper or on computer.

Data capture instrument
See Collection instrument.

Data catalogue
A facility for the storage of file specifications and locations.

Data collection
The process of gathering information for statistical purposes.

Data dictionary
A facility for the storage of metadata.

Data item
A particular characteristic of units in a population which is measured or observed.

Data management
The management of data from acquisition and input, to processing, and then output and storage.

Data mart
A repository for data related to one collection, or a number of related collections which share metadata, rather than data for an entire organisation. It is usually a subset of a data warehouse.

Data matching
The combination of two or more data sets at the unit record level.

Data modification
A method of maintaining respondent confidentiality in data by altering the identifiable data in a small way without affecting the

| | |
|---|---|
| | aggregate results. |
| Data reduction | A method of maintaining confidentiality of respondents in data by selecting appropriate aggregations or in presentation of the data. |
| Data retention policy | The rules governing the period of time for which data needs to be retained. |
| Data rounding | A means of maintaining confidentiality in tabular data by replacing small data values that could identify individual respondents with small random numbers. |
| Data suppression | A means of maintaining confidentiality in tabular data by not releasing information which may identify individuals. |
| Data warehouse | A central repository for all data within an organisation. A data warehouse is specifically structured so that numerous, wide-ranging combinations of data can be retrieved for analysis. |
| Demography | The study of human populations, their racial make-up, movements, birth rates, death rates, and other factors affecting the quality of life within them. |
| Demographic variable | A characteristic pertaining to the size, geographic distribution and density of human populations. Classic demographic variables include only age, sex, marital status, fertility, mortality and migration, but common usage has tended to include a wider variety of social characteristics such as: education, income, employment, etc. |
| Derivation | The process of creating new fields from data in existing fields, for example the derived item total income is the sum of all reported income items. |
| Derived variable | A variable created from two or more reported variables, for example total income is sum of all reported income items. |
| Descriptive statistics | Statistics which describe data (for example median, range, standard deviation) as opposed to inferential statistics, which are estimates made from the data. |
| Despatch and collection control system (DACC) | A means of tracking the progress of reporting units through the various stages of a statistical collection cycle. |
| Diary | A type of questionnaire in which respondents record for a specified period, their activities of interest to the survey e.g., purchases, T.V. viewing, trips made, as these take place. |
| Direct collection | The process of obtaining statistical information from the individual as opposed to obtaining the information from a third party as administrative by-product. |
| Dissemination | The distribution and communication of statistical information to users. |
| Dress rehearsal | A medium-scale quantitative test of the final survey design from start to finish, to test all aspects of the survey such as field operations, questionnaire design and processing. |
| Dwelling | A set of living quarters in which a person or group of persons reside or could reside. |

## E

| | |
|---|---|
| Econometric modelling | The use of statistics to model a real world situation, in order to test theories or make forecasts. |
| Edit failure | The act of a data item response not fitting the expected value for that item. |
| Editing | The process of checking and validating data. |
| Enumeration | The process of collecting data from the reporting units (individual members) of a survey population. |
| Estimate | An inference for the target population, using information obtained |

from a sample of the population.

**F**

Factor analysis — An analysis method which aims to explain the variation of a number of characteristics in terms of the relationships between a much smaller number of unobserved variables (factors).

Filter question — A type of question used to exclude a respondent from a subsequent question (or series of questions) if that question or series of questions does not apply to the particular respondent.

Fitness for purpose — The suitability of data for the intended use, that is, the degree to which the statistical information meets the data need.

Follow-up — One or more additional attempts made (in person, by telephone, letter, Facsimile, etc.) to contact a designated respondent where the initial attempt did not produce a completed interview or questionnaire.

Frame — A list of all members of a target population (for example people, households, businesses) which are available for survey selection.

Framework — A systematic and rigorous way of thinking about an area of interest, and promoting standards, consistency and comparability across data collections. A framework defines the scope of enquiry, delineates important concepts and organises them into a logical structure, showing the key relationships, processes or flows that exist between elements.

**G**

Geocoding — The process of identifying a dwelling, business location or agricultural location by a grid system, such as latitude and longitude.

**H**

Hypothesis — An assertion about a population, based on evidence, which serves as a starting point for investigation.

**I**

Imputation — The replacement of either missing or invalid data with accepted data in accordance with predetermined decision rules.

Index — A combination of weighted individual indicators, used to measure change without giving an actual numerical value. An index is usually given as a percentage, with the base period set to 100.

Inference — A conclusion about a population based on attributes of a sample.

Information development plan — A systematic means of cataloguing and addressing priority needs for data in a particular field.

Input data item — A particular characteristic of units in a population which is measured or observed and which is obtained from the respondent.

Input editing — The checking of individual responses to a survey prior to aggregation (before, during or after data entry).

Integration — The structuring of data to enable it to be used beyond the immediate purpose for which it was produced.

Intermediate edit — An edit which is changed or developed during the course of the collection.

Intermediate editing — The comparison of related units or forms to ensure consistency between individual responses.

Interpretability — A quality measure of the degree to which statistical information can be understood, explained and used.

**L**

Linear regression — A mathematical equation showing how one independant and one dependant variable for which the relationship between the variables is aapproximated by a straight line.

Logical edit — A check to ensure that two or more categorical items in a record do not have contradictory values, for example a respondent being 16 and receiving the age pension.

Longitudinal dataset A set of statistical data which observes the same analysis units over a substantial period of time.

**M**

Macrodata Aggregated and confidentialised statistical data.

Mark-in The notation in a despatch and collection control system to indicate that a survey response has been received.

Matrix A table of data.

Mean The arithmetic average of a set of values.

Median The middle value of a set of values that have been sorted in order.

Metadata Information about statistical data, for example contact person, how to access, accuracy, time period covered, scope, definitions.

Methodology The statistical theory (logic, principles and procedures) used to devise and conduct a statistical collection. See Collection methodology, Sampling methodology.

Microdata Statistical data of the smallest level. See unit record.

Missing data edit A check that data that should have been reported were in fact reported .

Mode The value or category which occurs most frequently.

Model An abstract mathematical problem that approximately corresponds to the real world problem.

Modelling The process of drawing together several variables and data sources to make inferences about the relationships between the variables.

Multi-stage design A sampling design in which a sample is selected in two or more successive stages. For example, the first stage could be a selection of electoral subdivisions, the second stage the selection of blocks of houses within these electoral subdivisions, and the third stage the selection of houses within the selected blocks.

**N**

Non-respondent An individual, or representative of an organisation, who does not provide information when requested in a survey.

Non-response bias Bias introduced to survey results if non-respondents possess different characteristics from respondents. See Bias.

Non-sampling error Errors that occur in producing statistical information that are not caused by sampling methodology. For example, errors can occur from the respondent, questionnaire, interviewer, processing, editing, field procedures, frame under-coverage, et cetera.

**O**

Omnibus survey A survey, initiated and conducted by a survey research organization at specific times or stated intervals, which contains several sections. Each section is customer-designed to meet the needs of an individual client. Each client receives the results of his/her section of the survey on an exclusive and confidential basis. Also referred to as "shared-cost", "multi-client" or "co-op" studies.

Optical character recognition (OCR) A form of imaging where the hand written responses on a survey form are extracted and converted to a useable form. The process involves capturing the electronic image of the form, then converting in into a useable format via repair and interpretation processes.

Optical mark recognition (OMR) A form of imaging where the response marks on a survey form are read and interpreted to give useable data.

| | |
|---|---|
| Outlier | An unusual value that is correctly reported but is not typical of the rest of the population. |
| Output data item | A particular characteristic of units in a population which has been collected and processed and is intended for dissemination. It may be an aggregate of data items. |
| Output editing | The checks applied to aggregate data once sampling weights have been applied. |
| Outputs | Reports, graphs, tables, publications, et cetera. displaying the aggregate data from one or more collections. |
| Overcoverage | The set of units which are on the frame of a collection but which are not in the scope. |

## P

| | |
|---|---|
| Parallel processing | Two forms of statistical activity on the same subject matter with the same objectives which occur simultaneously for the purposes of comparison. |
| Parameter | The value of a characteristic of a complete population. |
| Parametric item | A quantitative data item which has both an associated unit of quantity (for example dollar, hectares, hours) and an associated type (for example flow, stock, index, movement). See Classificatory item |
| Percentile | A numerical measure that also locates values of interest in a data set. |
| Pilot test | A pilot test is a small scale quantitative test of a survey, used to evaluate one or more of its components, such as the questionnaire or processing. |
| Poll | The questioning of persons selected at random or by quota to obtain opinions on a specific issue(s) or topic(s) which is (are) of current general interest. |
| Population | All the individuals or groups about which information is required, that is, the complete set of objects of interest in a statistical collection. A population may share a common set of characteristics. |
| Post-enumeration | The time in the process of a statistical collection after the data have been collected. |
| Post-enumeration survey | A study of a sample of respondents and non-respondents after information has been collected, with the aim of evaluating the quality of the data. This can occur after a pilot test or after the final survey. It questions how the respondent completed the form and gauges comprehension of survey concepts. |
| Principal component analysis | A technique for replacing original variables with a smaller number of uncorrelated variables, each of which is a linear combination of the original variables, so that the bulk of the variation can be accounted for using just a few explanatory variables. |
| Processing | The systematic transformation of collected data into statistical information such as tables and graphs. |

## Q

| | |
|---|---|
| Qualitative | Of interpretive or descriptive information based on opinions, perceptions, feelings and beliefs. |
| Quality | The standard of an output, assessed in terms of relevance, accuracy, timeliness, accessibility, interpretability and coherence and described in information (metadata) accompanying the output. |
| Quality assessment | A judgement on the quality of a data source with respect to a specific data need. It may be derived from a Quality Declaration. |
| Quality declaration | A statement about a data source, addressing dimensions of the data quality framework. |
| Quality framework | A template of the measures of quality, which is used to assess overall |

fitness for purpose.

Quantitative Of information based on numerical data

**R**

Range The difference between the largest and the smallest value.

Raw data Data as provided by the respondent, before editing, weighting and aggregation.

Rating scale A type of survey question (or set of questions) designed to record the direction and strength of a respondent's attitude(s) toward a specified topic or topics. The best-known types of rating scales are: Thurstone; Likert; Guttman; semantic differential.

Regression A technique for predicting the value of a variable from values of one or more other variables.

Relationship The way in which two or more characteristics are connected.

Relative Standard Error A measure of accuracy of a survey estimate, formed by the ratio of standard error to the estimate and often expressed as a percentage.

Relevance A quality measure relating to the pertinence of statistical output to the original objectives or the user's objectives.

Reliability The extent to which a survey would produce the same results, if repeated using another statistically equivalent sample and methodology.

Reporting unit The unit which provides the information about the unit of analysis. This mostly corresponds to the sampling unit. It may also be called a collection unit or a responding unit. See Unit.

Respondent An individual, or representative of an organisation, who provides information when requested in a survey.

Respondent load The effort, in terms of time and cost, required for respondents to provide satisfactory answers to a statistical collection.

Response rate The percentage of the target sample or population from which responses or usable data were obtained. The response rate can apply to the whole survey form or to the individual questions.

Robustness The ability of a statistic, system or methodology to cope with changes in the environment.

**S**

Sample The part of a population which is selected in a survey, for the purpose of studying characteristics of the entire population of interest.

Sample design A set of specifications which describes the population, frame, survey units, sample size and sample selection method for a particular survey.

Sampling error The error which arises because the data are collected from a part, rather than the whole, of the population.

Sampling methodology The principles and procedures involved in selecting a representative subset of a population for the purpose of drawing conclusions about the population. See Methodology.

Sampling unit A unit which is selected in a sample survey. Data from the sampling unit is multiplied by the sample weight to represent other similar, non-selected units. See Unit.

Scope The definition of the population of units about which information is required, for example households with children, retail businesses, people over the age of 65.

Seasonal adjustment The removal of the estimated effects of normal seasonal variation from a time series so that the effects of other influences can be more clearly recognised.

Self-enumeration
A method of data collection in which respondents complete the survey questionnaires without the involvement of an interviewer, for example in a mail-out survey.

Sensitive cell
A cell of a table containing a value which could lead to the identification of a respondent.

Sensitivity
(1) The degree to which data could change as a response to changes in assumptions or methods used in compilation. (2) The propensity of respondents to react to questions on a personal topic in a way that may adversely affect data quality.

Sensitivity analysis
(1) A method of analysing the sensitivity of a decision to a change in one or more of the assumptions used in making it. (2) An investigation of the degree to which the statistical data are affected by a change in the value of some parameter or variable, or by a combination of changes.

Significance-based editing
An editing approach which incorporates survey weights and estimation methodology into edits and maintains a link between individual responses and output estimates. It enables editing effort to be focused on units most likely to have a significant impact on final collection estimates. It can be performed at both the input and output editing stages.

Significance testing
The comparison of a quantity computed from data samples with theoretical values of standard probability distributions. Formally it is a comparison between a null hypothesis, H0 (for example that there is no difference between the means of two populations), and an alternative hypothesis, H1, (that a real difference exists).

Simple random sampling
A method of sampling where each member of a sampling frame has an equal chance of selection and each possible sample of a given size has an equal chance of being selected.

Stakeholder
Any individual or group with an interest in a particular collection. The interest may relate to the conduct of the collection or to its objectives and outcomes.

Standard
An accepted rule or definition for data items, collections, classifications, et cetera.

Standard deviation
The positive square root of the average of the squared differences of a set of measurements from their mean.

Standard error
A measure of the variation among the estimates from all possible samples, and thus a measure of the precision with which an estimate from a particular sample approximates the average results of all possible samples. The unit of measurement for the standard error is the same as the variable of interest.

Standard question wording
The wording of classificatory questions in a consistent way. For example, many survey organizations adopt a standard wording for demographic questions in all surveys, to reduce variations in answers which might be caused by different wording from one survey to another. It also aids comparability between surveys.

Statistic
The summary value of a variable or attribute calculated from sample data.

Statistical activity
(1) The designing or running of a census, survey or administrative data system. (2) Extracting or using data from a census, survey or administrative data system.

Statistical Clearing House (SCH)
An independent body which is the central clearance point for all Australian Commonwealth Government surveys involving 50 or more

businesses.

Statistical collection
Self-contained statistical activity involving the gathering, processing and combining of data on a particular theme.

Statistical cycle
The sequential process of gathering, analysing and disseminating information on a particular theme.

Stratification
The division of a population into strata, that is, homogeneous and mutually exclusive groups. Sample design and unit selection can be independently applied to each stratum, thereby increasing the overall efficiency of the survey.

Stratum
A sub-group of a population which is relatively homogeneous in relation to some known characteristic or set of characteristics.

Subject matter
The theme or area of interest for a particular survey or collection.

Summary measures
Measures of location (for example mean), spread (for example range), aggregation, et cetera of a set of data.

Survey
The collection of information about characteristics of interest from some, or all, units of a population using well-defined concepts, methods and procedures.

Survival analysis
The analysis of the association of variables with the time to an event.

Synthetic estimation
The use of a known value of one variable to predict the value of another.

### T

Tabulation
A counting of the number of cases falling into each of a number of categories.

Time series
A statistical record of a particular activity where the data is measured at regular intervals over a period of time, for example monthly unemployment rate. Time series are collected on this basis to assist understanding of the current situation, enabling the most recent data observations to be placed in a meaningful historical perspective.

Timeliness
A quality measure relating to (1) the time taken between the occurrence of the characteristics/events being measured and the release of statistical output and (2) whether the output of a collection is sufficiently up-to-date for the user's purpose.

### U

Unit
The entity used in the design, collection, compilation, tabulation or publication of statistical data. There are many types of unit. See Analysis unit, Reporting unit, Sampling unit.

Unit record file
A list of data relating to individual population members.

Univariate
Of a single variable.

### V

Validation
The checking of data values to ensure consistency either within the data source or with data from different sources.

Validation edit
A check of the validity of a basic identification or classificatory item in unit data, for example sex is coded only as one of M or F.

Variable
A characteristic that may assume more than one of a set of values to which a numerical measure can be assigned (for example income, age,

weight, et cetera.).

Variance — The arithmetic mean of the squared deviations from the population mean.

**W**

Weight — The number of units in a population represented by a particular unit in a sample (for example a weight of 20 means that the sampled unit represents 20 units in the population).

Weighting — The procedure for applying weights to survey results