

3 Data development principles and methodology

The first part of this chapter outlines the main principles that underpin good data development. These principles are applicable regardless of the purpose of data development, such as statistical purposes or electronic records, and are not listed in any order of importance.

The second part of the chapter describes a sequential, but overlapping, approach to data development. It outlines the steps involved in data development from the time a need for data development is identified to the endorsement of a fully developed data set.

3.1 Principles of good data development

3.1.1 Creating data standards is part of data development

The quality of data, including its consistency and comparability, is enhanced when data standards are available to support the collection and use of a data set. The development of data standards is not something that is done at the end of the data development process. It is very much a part of the data development process.

Data developers must develop operational procedures to ensure that data standards become a key component of the data development process, and to ensure their ongoing relevance and maintenance.

There is a cost associated with creating data standards, but the cost of not creating data standards is likely to be even higher. This includes the loss of information that occurs due to staff changes, data redundancy, data conflicts, liability, misapplications and decisions based upon poorly documented data. These costs should be factored into the data development budget.

3.1.2 National and international standards should be used wherever available and applicable

When data are being developed, it is important to ensure that the specifications of the data are consistent with national and international data standards. This avoids duplication of effort and the development of conflicting data standards.

Data sets should be based on a single set of agreed definitions and data standards, such as the NHDD, so data developers can 'pick and mix' from existing agreed definitions, ensuring a high degree of consistency and reducing data development time and cost.

Local needs sometimes require more specificity than national or international data standards can provide. Where this is the case, it is possible to specify more granular data standards as long as they are mappable to national or international data standards. Tips and examples on what to do when local requirements demand more detailed or granular data are provided

later in this chapter in the section entitled 'What to do when a modification to an existing data element is required' and in the following chapter in the section entitled 'Value Domain'.

3.1.3 Be clear about the purpose of the data collection

Before deciding on what data to collect and develop, it is important to be clear about the purpose of the data collection including the important policy or planning questions, or service provision needs for which data collection is required, and how having the data will help deliver more efficient and effective services. It can be tempting to decide on what data to include in the collection before the purpose of the data collection is fully defined.

3.1.4 Data included must be required to meet the objectives of the data collection

Once the purpose of the data collection has been defined, the data required to meet the objectives of the collection can be identified and developed. Data developed and included in the data collection must meet the collection's objectives as the cost of developing, collecting, compiling, validating and reporting data can be expensive.

3.1.5 Create once, use often

The availability of existing data sources should be explored and used where possible.

An important principle of data development is that data needed to support secondary (or downstream) information purposes (such as reporting, policy, governance or decision support) should be derivable from primary data (point-of-service delivery data). Otherwise, data needed for downstream requirements would have to be developed and collected separately, resulting in significant additional costs because of the need to establish parallel data collection systems to support existing and new data flows, whose products may not necessarily be fully compatible. Similarly, data developed and collected for mainly statistical purposes should be used to provide feedback to improve and enhance primary service delivery.

3.1.6 Acknowledge the limitations of data

While it is important to be mindful of the opportunity to re-use existing data, it is also important to be aware of any limitations of the data and to ensure that these are acknowledged. For example, measures such as child abuse can be sensitive to variations to practice over time, such that a similar case may be substantiated over one period of time but not in another, making comparisons over time problematic.

It is also important to note that data that are suitable for some purposes may have limited use for other purposes. For example, while data about 'presenting problem' and 'reason for encounter' may be useful measures of hospital emergency department services for treatment of patients and for management of department resources, they may not be adequate to inform about the incidence of domestic violence or child abuse.

3.1.7 Data development may be incremental

Data development should support incremental development, such that the scope of the data set is expanded over time.

It may not be possible to develop all data required for a data set at the same time. Some data may be more readily agreed upon and easily collected, while other data may be more problematic and require more time to develop. For example, for a problem gambling support services data collection, data developers may find it easier to develop demographic data about clients seeking help in relation to their gambling problems, but more difficult to develop and agree on data that can be used to measure outcomes of problem gambling educational programs or assistance services. Depending on the timelines and resources available, it may be better to stage the data development such that the scope of the data collection is expanded over time.

3.1.8 Data development is system independent

Data development must ensure that data in the data set is well defined and standardised so that it can be compared independent of the organisation, system or tool that captures the data.

Data development must not be limited by the capability of any particular system.

3.1.9 Data development should be mindful of privacy concerns

Data development processes must take into account security policies and privacy issues, including ensuring compliance with the information privacy principles. Data sets should avoid the inclusion of data that may be regarded as private or confidential in nature. Otherwise data provided may not be reliable or accurate because respondents may be reluctant to provide the data.

3.1.10 Data development should minimise collector/recording burden

Good data development should ensure compatibility of data collection and reporting requirements to avoid situations where the same data has to be collected, counted or reported differently for different programs. This will reduce the reporting burden on service providers and help to reduce cost..

3.1.11 Data development should reflect not drive practice

Data must be able to be collected as a by-product of service delivery or administrative practice. The data must be relevant and meaningful to those collecting the data and be of benefit to service providers. Data development must take into account the business needs, feasibility of data collection, appropriateness of the data, as well as scientific evidence and recommendations (guidelines) of subject matter experts. Where possible, data development should be based on data that service providers already want or need to collect about clients and service provision. Most importantly, data development should ensure that data collectors are not constrained or forced to operate in ways outside usual practice. For example, data about the country of birth of a client is usually not relevant for service delivery

and should not be part of routine data set collections. Of more relevance to service delivery is data about whether interpreter services are required, client's preferred language, ethnicity and religion. Data about country of birth, which may be useful for statistical purposes, can be collected as part of a 'one-off' survey of clients.

3.2 The data development process

Data development is a methodological process, based on an understanding of the information to be derived using the data. It includes modelling data needs and clarifying the relationships between data. Key data concepts are identified and standardised using data elements. Data development results in the production of a set of data standards to ensure consistent collection and use of the data set.

3.2.1 Business context and information need identification

It is important to be clear about the purpose of the data development project and the benefits it has to offer, otherwise the effectiveness of the resulting data set may be greatly reduced. It is therefore necessary to obtain an understanding of the business context within which the information is needed, and the high priority policy issues, questions and/or service needs before proceeding with data development. For example, where the objective is to ensure equitable health care to non-English-speaking populations, an important policy question could be: 'Do we provide an equitable level of health services to people from non-English speaking backgrounds?'

Needs analysis results in documentation of the business requirements for data (or a problem statement), the target population (for example, healthcare professionals) and service environment, a description of the problem, gaps between expected and ideal outcomes, and the relative priorities of the business requirements. This helps to guide further data development, including identification of what data are required.

While the need for data development may relate to a specific problem or purpose, the expected benefits at all levels (for example, policy, program development, performance management and service provision levels) should be acknowledged and clearly stated.

3.2.2 Feasibility analysis

Developing a data set can be costly. Therefore, it is important to undertake feasibility analysis in order to provide an indication of the scope of the data development and the resources required to support the project. Feasibility analysis includes the identification of stakeholders and their relevant level of interest, the data needed to support information requirements, analysis of what data are currently collected, and if sufficient resources are available to proceed with the data development project.

Identifying stakeholders

For the data development process to be successful, it must reflect the shared interests of those who use or rely on the data. The feasibility study should identify relevant stakeholders, their information requirements, their availability to participate in data development activities, and their level of commitment to data development. Stakeholders

could be service providers, service managers, national, state and territory governments or researchers.

What data currently exists?

It is important to identify existing data sources and assess their currency and usefulness for the current purpose. It is not always possible to get information about every system used for collection purposes, but obtaining a representative sample is recommended. The feasibility study should identify information gaps and data that are/might be problematic in terms of obtaining accurate information.

Where possible, data should be existing by-products of the service – that is, service providers already want or need to collect the data. When developing data for secondary uses (in relation to answering key policy questions) it may be necessary to undertake a cost-benefit analysis to determine whether it is cost-effective to collect data that are not directly required to meet service delivery needs.

As part of the feasibility study, it would be useful to assess whether the data that are currently being collected allows for comparison over time and if there are any existing standards, performance indicators or benchmarks.

How the data are currently collected

An understanding of the collection methods and how data are currently being reported assists data development. Data development that is mindful of the operational constraints of the systems and people who need to record the data results in a more usable and reliable data set. Data collection methods may include paper, various electronic media, face-to-face or telephone contact. Data collection is normally in the form of either an ongoing or snapshot collection.

Reporting arrangements and periods

Reporting arrangements should specify if there are any current reporting requirements for existing data. Reporting arrangements usually vary between organisations/jurisdictions and data developers need to be aware of the differences and challenges.

Future reporting requirements need to be established. In the case of NMDS reporting, there is normally a two-stage process – from service providers to jurisdictions and from jurisdictions to a national collection agency. Jurisdictions are provided with a national specification and agree to collect and transmit the data in that form. Some collections can be reported directly from service providers to national collection agencies.

Scope

The scope of the data set should at least specify:

- the population that is the target of data development
- the services, agencies or organisations to be included and excluded
- information to be included in the data set (such as interventions, conditions or services).

The scope of data development may initially be limited depending on whether the data required already exists and the time that is available for data development. However, future expansion of the data set to include other priority data should be considered and planned.

This would enable data that are currently difficult to collect or which requires a longer time to develop, to be included.

Statistical Unit

Where the purpose of the data collection is for statistical analysis, it is important to give some consideration to the statistical unit by which the data will be grouped. The statistical unit determines the level at which data will be collected and/or reported and is fundamental to a data set. For example, for a client-based collection the data could be collected at the client level, at the service contact level, the service episode level or at the service provider level (all clients to which a provider has provided services). A decision also needs to be made about whether the data needs to be aggregated or reported for each client.

Timeframe

Based on an understanding of the expected information requirements and current data availability, the study should provide an estimate of the time required to develop the necessary data.

Resources

The feasibility study should provide an indication of resources that will be required to undertake the development of the data set. Costs include salaries, project management fees, costs associated with data set and data standards development, consultations with stakeholders, endorsement of data standards, travel and associated costs, meetings costs, stationery, publication fees and pilot testing costs.

At this stage, it may not be possible to provide an indication of the cost of implementing the data set as more detailed analysis of the data for development by the data working group will need to occur.

Establish a project team and data working group

Once feasibility has been accepted and a decision made to proceed with data development, a project manager (or team) is recruited to lead and manage the data development. It is worthwhile forming a dedicated project team to lead and manage the data development project, rather than trying to undertake data development as part of an ongoing committee's work. This is because a data development project is resource intensive and requires considerable time and effort to lead, coordinate and promote. One of the first tasks of the project team is establish a data working group, if one does not already exist. The data working group should bring together stakeholders, who may include service users, service providers, policy makers, data experts and other key advocates. The data working group is usually established under the auspices of an authoritative body in the field which has recognised the need for, or is financially supporting, the data development work.

The role of the project team is to manage the data working group in relation to the data development project and liaise with the patron or funding body responsible for the project. The project team not only leads the data development work, but also promotes and coordinates other related activities.

In addition to the working group, it may also be necessary to establish a reference group to seek specialist subject-area advice and feedback throughout the data development process. The reference group, while providing valuable advice and feedback, is not expected to do the

actual data development work. Reference group members include stakeholders in the area for which data are being developed. A range of expertise is important to ensure that a robust data set that is useful from both a data collection perspective (primary use) and a policy perspective (secondary use) is produced.

Some of the tasks for the project team and the data working group would include:

- developing a project plan or work program
- collaborating and consulting with stakeholders
- identifying core data that needs to be collected in collaboration with stakeholders
- leading technical discussions on the data development
- developing data standards using data elements to support the data set
- circulating draft data specifications for public comment
- considering views expressed and comments made
- organising field tests and analysing data and feedback from the tests
- estimating cost of implementing the data set
- liaising with data committees regarding endorsement of the data set.

3.2.3 Consultation and collaboration

Consultation with stakeholders enables all interested parties to have input into the data development. Requirements or comments provided inform the final decisions to be made about the data set.

Consultation should occur from an early stage and continue throughout the data development process. Interested parties then become acquainted with the work and make contributions towards it, while amendments can still be made and comments taken into account. More importantly, consultation facilitates the uptake and dissemination of the data standards. During consultation with stakeholders, agreement is reached about the data elements, including the meaning of the data elements and how they are to be represented and collected, to ensure consistency and comparability across organisations.

If the data collection is of national significance, and is to become part of the existing national health information system, it will need to be endorsed by the key national committees concerned with obtaining agreement on national health information (see Chapter 6). The Health Data Standards Committee (HDSC) and the Statistical Information Management Committee (SIMC) should be informed of the project and given an opportunity to provide input and/or advice. These committees should be regularly consulted and updated regarding the progress of the data development project.

Procedures should be established so that adequate time and opportunities are provided for written comments. Essential information in relation to the data development project should be easily accessible to all interested parties. Openness, democracy and involvement in this process by all stakeholders are important.

Combined Top-down / Bottom-up approach

Whether data development is policy driven (top-down) or initiated by service providers (bottom-up), the data working group needs to adopt a combined top-down/bottom-up approach in order to identify data that are to be developed and included in the data set.

Data development that is initiated by policy requirements needs to be mindful of the operational constraints of the systems and people who need to record and process the data. These constraints are particularly important when the data needed for management reporting is a by-product of service delivery.

Similarly data development that is initiated by service providers, or those who will be collecting the data, should consider the statistical impact of the data and how it can be used to measure service performance.

Top-down approach

Data development that is initiated from the top down is usually in response to information required to inform policy initiatives, service utilisation, program outcomes, to support planning decisions or measure performance, or to inform about cost of service provision. The data that are required to support these information needs are usually aggregated data, such as cost of service, staffing requirements, number of clients requiring service or total sales. A major challenge for data development is to ensure that data used within the service delivery setting can be accurately and meaningfully aggregated to produce statistics that inform policy and decision making.

The top-down approach leads to an understanding of how the required information will be used, such as in benchmarking or governance, and will provide guidance in relation to what we want to know or measure, the data needed to be able to achieve the measures, how the data should be represented and the corporate reports that would include these data.

Bottom-up approach

Data development that is initiated from the bottom-up is usually in response to information required to assist service providers to provide high-quality services more quickly and efficiently. For example, this would include data that are required to alert service providers when clients are due for services, or when services cannot be provided to clients for specific reasons, data for decision-support systems or data to facilitate electronic delivery of services.

This type of data usually has more detail about services provided and clients (for example, data that uniquely identifies clients and their service needs).

The bottom-up approach involves consultation with service providers who are directly involved in the collection and recording of data and have an understanding of the day-to-day issues, and experience with use of the data that are collected. The bottom-up approach provides information about the data collection and maintenance within the local or service delivery setting and leads to an understanding of the context and feasibility of data capture. It is through consultation with service providers that the data working group becomes familiar with data that are collected or required, how these data are collected and used, the terminology used, and if any difficulties are experienced in collecting any data.

A combined top-down/bottom-up approach can ensure reliable integration of requirements at service delivery and policy levels and development of data that can be appropriately aggregated for different purposes.

The top-down/bottom-up approach resolves the tensions that often exist between information that is useful for evaluation at the policy level and ease of collection at the service delivery level. Data that are easily collectable may not be comprehensive enough to undertake an evaluation of service provision and direction for policy and planning. On the

other hand, if data are difficult to collect, or not perceived to be useful at the service delivery level, then it becomes time-consuming and cumbersome.

In general, if management or policy data are captured via routine administrative processes that do not require the data for the main job at hand (such as processing a payment) then data quality may be at risk.

3.2.4 Identifying data for development

An important part of data development is the detailed analysis of the data that are necessary to support business requirements, which, in turn, depends on the purpose of the data development, the sorts of questions that need to be answered or problem that needs to be addressed. Core concepts are identified and these are defined and standardised. For example, if equitable access to problem gambling support services is a focus area, then information should give an indication of equitable access to these support services to various population groups. There should be a decision made by the data working group about how to measure equitable access. For example, equitable access could be measured by the percentage of people in a population group accessing a service or the availability of interpreter services, and so on. In this way, the data to be collected are identified and each data identified is described and standardised using a data element.

As some of the required data may be less readily available, or difficult to collect consistently (such as the amount of money spent on gambling), and there is a cost and effort involved in data collection, it may be necessary for the data working group to identify a core (or minimum) set of data for collection, based on how the data can be collected in practical terms and the priority of the data. Future expansion of the data set to include data that requires a longer time to develop would set up a work program for future data development work.

Modelling data requirements

A data model is a useful tool when identifying and depicting data for development. The model provides a diagrammatic representation of the building blocks of data necessary to meet user information requirements. The model is used to provide the framework and context for the data identified and described. It is employed to help users identify and articulate data requirements and specify the business rules and relationships that exist between data. It is important that the data model is agreed to, and signed off, by stakeholders before proceeding with data development. Otherwise, the data working group may find that they have to deal with changing data requirements throughout the data development process.

In a data model the 'thing' of interest is called an entity and data elements are used to describe 'things'. For example, a data element 'date of birth' would be used alongside other data elements to fully describe a 'client'. In this case the client is the thing of interest, which in a data model is depicted as an entity. Each entity needs to be defined and described through its associated data elements. For a specific purpose, a client may be fully described by date of birth, sex, postcode, interpreter service required and preferred language. The data model depicts all the entities and data elements that are required to meet information requirements. Identifying the data elements that are required, and agreeing to their definitions and value domains, is of critical significance to consistent data collection and use.

An example of a simple data model depicting entities and their relationships is shown in Figure 2. Entity-relationship models are useful for depicting high-level relationships amongst entities.

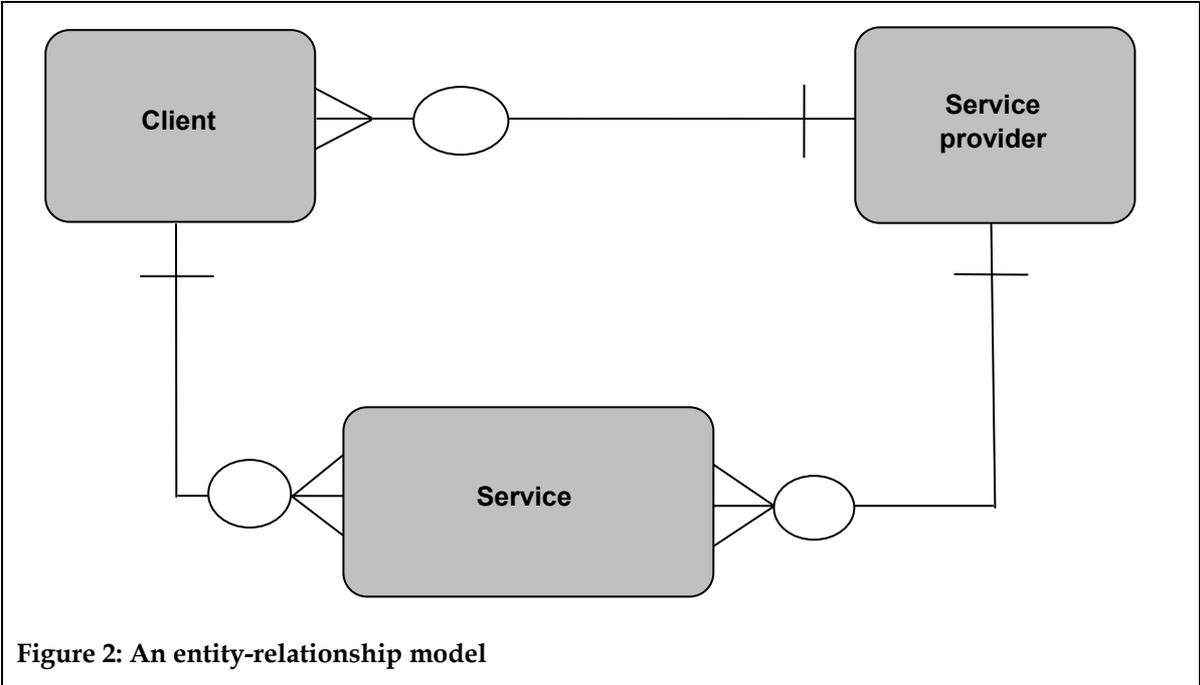


Figure 2: An entity-relationship model

Entities

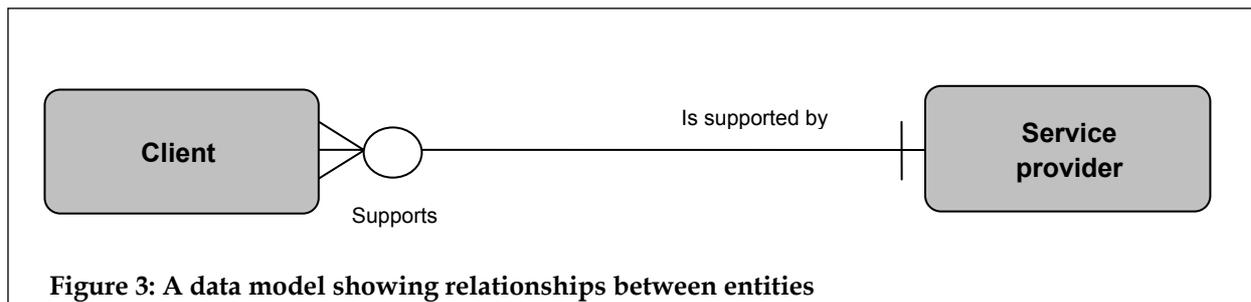
The boxes in the model represent entities. The identification and definition of entities is an important part of any information.

When questions are asked such as ‘How many clients receive support services?’ this assumes a shared understanding of the term ‘client’. Other questions such as ‘What types of support services were required by clients?’ suggests other entities are required, such as ‘services’ and ‘service provider’.

The definition of entities is very important, particularly for consistent reporting, because the definition should help to standardise what is included, and what is excluded, within the scope of an entity.

Relationships

Entities have relationships (or rules of association) with other entities in the data model. The lines and arrows in the model represent relationships. There are many different conventions that describe the meaning of the relationships between entities. In Figure 3, the line into the entity ‘Service provider’ is an example of an optional, one-to-many relationship which can be interpreted as ‘A client must be supported by one (and only one) service provider. The line in the other direction can be read as ‘A service provider may support one or more clients’. The ‘O’ means the relationship is optional while the vertical line indicates that the relationship is mandatory. In this instance, it is assumed that a service provider can exist without a client, but a client cannot exist without a service provider.



Decisions on relationships and consistent implementation of relationship rules are fundamental to consistent and effective informational collection and use.

Data elements

Once the essential entities have been defined, the characteristics that describe the entities are identified and defined. These characteristics are turned into data elements and are used to standardise the agreed meaning and acceptable representation of data for use within a defined context.

It is important that each data element describes only one concept and therefore it may be necessary to create more than one data element to describe the data that are required. Since there is a cost associated with data collection, every data element should have a clearly defined purpose.

The formulation of data elements is described in more detail in Chapter 4, while the next section describes when to create new data elements, when and how to modify existing data elements and when to reuse existing data elements.

3.2.5 Developing data elements

Re-using existing data elements

In some instances, the data elements that are required for a data set may already exist. Wherever possible, data developers should identify opportunities to use existing data elements from the authoritative sources, such as the National Data Dictionaries, ABS catalogue and Standards Australia. This saves unnecessary effort involved in recreating data elements and promotes alignment with existing standards. If an existing data element adequately describes the concept to be collected, then this data element should be used. However, data developers should not be limited to re-using an existing data element if it is not relevant or is too constraining for the needs of the data set.

Modifying existing data elements

If there is close alignment with an existing data element, but it is not exactly what is required, there may be scope to modify the existing data element to meet new requirements. It may be possible either to modify an existing data element without affecting its current use or to include collection-specific information that is not inconsistent with the use of the current data element. Before changing an existing data element, the data developer should consult with others using the data element.

In deciding whether an existing data element can be used or if it can be modified, the data developer needs to ensure that the existing data element is conceptually the same. Even if it is conceptually the same, the concept may still need to be represented differently for a different purpose. In this case, a new data element is created, instead of re-using or modifying the existing data element. For example, if the pre-existing data element 'height of person' specifies measurement in metres and the current data set is about newborns, it may be necessary to create a new data element that specifies measurement in centimetres. In this case, the two data elements are the same conceptually, but because the concept is represented differently, separate data elements are required.

What to do when a modification to an existing data element is required.

If it is established that an existing data element could be used, but only with modifications, the first thing that should be done is to identify users of the existing data element. The steward of the data element may be able to provide information about current users of the data element or, if not, data working group members and stakeholders of the project are often able to assist. The project team will need to communicate with existing users (for example, by attending data working group meetings of the existing user) to collaborate on a joint solution that would be workable for existing and new users of the data element. This can sometimes be a slow and time-consuming process because it may take time for existing users to assess the impact of the change. It is therefore important to establish the priority of the data and the importance of the modifications before seeking changes to the data element.

It is important to determine if the existing data element currently conforms to agreed national, state or international standards, where they exist and how the modifications required will impact on compliance. Where modifications to the value domain are sought, it is important to ensure that it is consistent, or mappable, to existing data standards wherever possible.

It is often the case that data standards developed for national (or even state) purposes are of low specificity (less granular). This may not satisfy local needs where more detail is required. Similarly, clinical users of health data often require very granular data, while statistical users of health data are usually satisfied with less granular data. Depending on the purpose, value domains may be expressed to varying degrees of granularity as long as they are mappable to the recommended standard. A one-size-fits-all approach is not necessarily required to achieve standardisation. For example, the national standard for eye colour may include only:

1	blue
2	green
3	grey
4	brown
5	black
6	other

Locally a broader range of colours may be required. The additional range must be mappable to the national standard. For example, dark-green and light-green are mappable to 2-green (many to one), or dark-blue is mappable to 1-blue (one to one). A value such as greenish-grey is not an acceptable addition since it can be mapped either to 2-green or 3-grey (one to

many). Such a case should ideally be taken to the appropriate governance body responsible for the national standard for discussion before being adopted at the local level. Similarly, local additional values must be able to be mappable to 'other' in the national standard, or to one of the other existing values as described in the previous example. More examples of compliance and mapping to existing standards are provided in the 'Value Domain' section in the next chapter.

Creating new data elements

Creating new data elements includes formulating and defining the data element and specifying associated rules for when and how the data are collected. It may be possible to re-use some of the components of existing data elements, such as existing value domains, or data element concepts for the new data element where relevant. The following chapter outlines how to create data elements.

3.2.6 Field testing

Having developed a draft set of data elements, the next step is to field test the data set. Normally field testing is conducted for the whole data set rather than for individual data elements.

The purpose of field testing, or pilot testing, is to:

- test the practicality and consistency of the interpretation of data elements and the comprehensiveness of the codes (range of possible answers)
- test the quality of reported data
- assess how easy, or how difficult, it is to collect the data
- test the transmission of data and data analysis
- identify guidelines for implementation that need to be added to ensure consistent interpretation and reporting.

A selection of service providers or agencies are identified and recruited to participate. It is desirable that the sample include a range of agency/service types, models of care and locations.

The pilot test is usually conducted over a period of approximately 2 to 4 weeks.

Pilot test participants must be able to provide feedback on the data set and the collection process, both verbally and in writing.

The outcome or findings from the field test should be used to refine the data set before it is implemented.

3.2.7 Estimating cost of implementation

There are costs associated with implementing a data set. In some instances, a business case may need to be submitted before a data set is approved for collection.

The business case must consider the following:

- The cost of the change or new data.

- The ability to make a 'generic' change to the data collection, or the collection process that would manage this change, and other related issues. If there is the ability to institute such a modification, this should be costed.
- The opportunity and timing relevant for data collection review. Where the information needs have changed, there should be a review of the whole process of information collection in order to identify the most relevant change. It is easy to simply change the data collection rules or values, but, over time, this can result in a disjointed and complex collection system. When considering the introduction of any change, it is important to consider whether it would be cheaper in the long term to restructure the data collection rather than to make further 'bit' changes. It is often significantly cheaper to recognise opportunities for change. For example, as systems move towards electronic health records, the need to develop new systems for data extraction will arise. This provides an opportunity to build table-driven extraction rules that can be readily maintained and updated without the need to change software, thus offering significant future flexibility and potential to reduce the costs of later changes.
- Cost of making no change.

Implementation costs will vary depending on a range of factors, including, the extent to which systems are already in place and data set requirements over and above current system capabilities and reporting arrangements (for example, whether the data are sent directly by data collectors to the collection agency or if it is first sent to an intermediate agency for checking and compilation). With national data set collections, data from service providers is often transmitted to a central point within state/territory jurisdictions, before it is forwarded onto the national collection agency.

The cost to data collectors (service providers) include any system changes, training and other requirements necessary to facilitate accurate collection, extraction and transmission of data. Actual costs will depend on decisions made about scope, reporting arrangements and individual data elements in the data set.

Where jurisdictions are involved in the checking and compilation of data from service providers, there are costs associated with receiving, checking and loading data from service providers, training costs to ensure new systems are used appropriately, staff costs (additional staff required to support collection, collation and data supply) and system costs. Where there is no intermediate agency, many of these costs are borne by the collection agency.

Collection agency costs include costs required to develop and maintain a system to receive, process and store data received from data collectors. There is normally an initial set-up cost, and an estimate for ongoing maintenance.

Each time a data set is expanded, an estimate of the additional costs of proposed data set changes should be prepared.

3.2.8 Obtaining authoritative endorsement

Before the data set can be implemented, it is important to obtain endorsement of the data set and its accompanying data standards from the appropriate authoritative body. This is important to ensure credibility and uptake of the data set. For example, where a data collection is to become a NMDS for use in the health sector, the data standards need approval by the HDSC and the NMDS needs to be approved by the Statistical Information Management Group (SIMC) before it can be endorsed by the National Health Information

Management Principal Committee (NHIMPC). Similarly, governance arrangements exist for the endorsement of data collections for use within the community service and housing sectors. Chapter 6 provides more detail about data agreements that exist at the national level within these sectors, and the processes in place to obtain the endorsement of national data collections and data standards.

Many state/territory departments have documented processes in place for the approval of new or modified data sets, and accompanying data standards, for use in local jurisdictions.

3.3 Risk management

There are a number of risks associated with developing data sets. Some of these risks are as follows:

- Changing information needs – as services grow and expand, and as service models and policies change, it is reasonable to expect that information requirements will also change. Accordingly, there should be mechanisms in place to ensure that data sets are periodically reviewed and improved in order to reflect the changing information needs. This ensures the continued relevance of data sets in answering questions that are of importance to relevant parties. Data that are no longer relevant may need to be dropped from the data set, while new data may need to be developed and included.
- Failure to re-use data – opportunities to re-use data that are already developed should be investigated in order to avoid costly duplication and reduce possible conflicts in data. It is important that data required for secondary purposes (for example, policy initiatives, planning and measuring performance) is obtained as a by-product of service delivery. For example, e-health information initiatives, which are primarily concerned with communication and sharing of information that is relevant to clinical care, provide an excellent opportunity for the development of new statistical collections that inform and compare service delivery and measure performance. Statistical impact analysis should be undertaken to maximise the scope for extraction of useful statistical information from data developed for mainly primary (or administrative) purposes.
- Poor-quality data collected – periodic analysis of the data collected can help to identify data quality issues. The presence of poor-quality data may be an indication that the data collected is not useful or has been poorly understood by data collectors. Training or education campaigns may be employed as a means of addressing and emphasising the value of good-quality data. However, where it is found that the data are not of value to actual service delivery, it would be worthwhile undertaking a cost-benefit analysis of the need to routinely collect the data as part of the data set.