Data Quality Framework

Justice Sector Information Strategy
Ministry of Justice
June 2008
Version 1.0

Contents

Document History

Date	Author	Summary of Changes	Version	Status
01/05/08	Mandy Mackay	Initial draft	0.1	Draft
30/06/08	Mandy Mackay	Incorporating feedback from DDWG	1.0	Final

Definitions

It is important to define the terms that clarify and provide the distinctions between data, information, and knowledge.

Knowledge differs from data or information in that new knowledge may be created from existing knowledge using logical inference. If information is data plus meaning, then knowledge is information plus processing.

Term	Definition
Data	Data is a set of discrete, objective facts about events. Data could be described as facts and figures without context and interpretation.
Information	Information is data that has been processed into a meaningful form and content relevant to a particular situation. Data is transformed into information by adding value through context, categorisation, calculations, corrections, and condensation. Information management is the planning, budgeting, control and exploitation of the information resources in an organisation. The term encompasses both the information itself and the related aspects such as personnel, finance, marketing, organisation, and technologies and systems.
Knowledge	Knowledge is the psychological result of perception and learning and reasoning. Knowledge management is the capability to create, maintain, enhance, and share intellectual capital across the organisation in support of business or sector objectives. It is achieved by developing a conscious strategy of getting the right knowledge to the right people at the right time and helping people share and put information into action in ways that strive to improve organisational performance.
Data Quality	Data that is fit for all the purposes justice sector agencies will use it for.

From previous work completed on data quality in the justice sector, fifteen data quality dimensions were identified to assess the quality of data, dependent on the context and use of the data. The justice sector has identified four data quality dimensions as being of the highest priority in the assessment of data quality, *Accuracy, Completeness, Timeliness,* and *Accessibility*. These dimensions are described in further detail in Appendix A.

The Data Quality Framework should be read in conjunction with the Data Quality Toolkit resources set out in Appendix B, to get a comprehensive view of managing data, information, and knowledge within the justice sector.

1. Background

The Justice Sector Information Strategy (JSIS) provides a framework for improved collaboration to manage and share relevant information between justice sector agencies from a collective perspective. Each agency within the sector requires accurate, relevant, and timely information to manage operations and to develop and review policy. To achieve this, data is shared between the justice sector agencies in accordance with government legislation.

The justice system is an information-rich, integrated, and highly complex environment. Information is shared across a secure virtual network of seven operational systems on a daily basis. Approximately 32 data exchange interfaces facilitate around 14 million transaction and event records annually.

Increased collaboration within the justice sector is necessary to provide better access to, and make better use of, a range of data and information resources and tools, to support strategic decision-making. To maximise the benefits to be gained at all levels of the sector, a high level of data quality is required.

The importance of using high quality data on which to advise Ministers, base strategic decisions, and continuously improve the justice sector business, must not be underestimated. Every employee of the justice sector is responsible for providing high quality data and has the right to quality data.

Concerns have previously been expressed across all levels of the justice sector at the need for improved data quality. Documentation on the causes and costs of poor data quality can be read in Appendices C and D. To date, there have been no metrics or benchmarking provided to highlight the problem areas. This prompts the question – are there definable data quality issues, or is the matter simply one of perception?

The purpose of this document is to provide a framework for the management of data quality across the justice sector. This will enable the question regarding the status of data quality to be answered ongoing. To understand what a Data Quality Framework is, it is necessary to first define what data quality is.

> *"Data are of high quality if they are fit for their intended uses in operations, decision-making, and planning. Data are fit for use if they are free of defects and possess desired features"* Dr. T. Redman – Data Quality: The Field Guide (2001).

Agency representatives on the Data Definitions Reference Group (DDWG) have agreed the following definition for data quality.

> *Data that is fit for all the purposes justice sector agencies will use it for.*

Simply put, it is having the right data, at the right time and place, to the right people for the right use.

Based on the definition of data quality, a data quality framework can be described as:

> *"At its most basic, a data quality framework is a tool for the assessment of data quality within an organisation (Wang and Strong, 1996). The framework can go beyond the individual elements of data quality assessment, becoming integrated within the processes of the organisation"* (Kerr, 2006).

The Data Quality Framework contributes to the following themes of the Justice Sector Information Strategy:

Theme 1 - Improve the quality and integrity of justice sector operational data assets
Theme 2 - Effectively manage shared justice sector data and information
Theme 3 - Ensuring we support strategic decision-making in the justice sector.

The Data Quality Framework is also strongly aligned with the Justice Sector Information Strategy vision. The vision at the core of the strategy is:

- We have a high quality, dependable, and valued information base that supports operational and strategic decision-making
- We continue to show leadership in information sharing and collaboration.

2. Problem Statement

The existing justice sector Data Quality Assessment (DQA) Methodology is a stand alone document. The Justice Sector Information Strategy (JSIS) provides for a data quality programme, which currently consists of a series of stand alone data quality assessments based on this DQA Methodology.

The DQA Methodology is used to assess the quality of data identified for detailed analysis. The DQA Methodology is a qualitative study of data and records a snapshot of the relevant data process at a set point in time. The result of applying the DQA Methodology is the identification of problems and possible recommendations.

A review of the DQA Methodology identified a number of weaknesses in assessing data quality. For example, recommendations are made from a data process perspective with no quantitative measures or goals. Data quality issues identified as a potential risk by this process may never be realised or eventuate. The DQA Methodology does not ensure resolution or follow through of recommendations, a key requirement to improve data quality.

This creates difficulties for the sector in determining the status of recommendations. There is no clear understanding of the impact of complementary items of work that may or may not supersede the data quality assessment recommendations. Recommendations may not necessarily be implemented. For those that are, there is no post-implementation review, and there is no measure of whether the initiative implemented provided the intended results.

Another issue with the existing DQA Methodology is that it does not allow for flexibility in the analysis and problem solving of data quality issues. It is accepted that there are some data quality issues which require urgent or immediate attention. In these instances it may not be practical to complete a full data quality assessment at that time or wait for a scheduled data quality assessment.

The scheduling of data quality assessments is currently based on perceived data quality issues. Assessments to be completed are approved by the Justice Sector Information Strategy Management Committee (JSMC) as part of annual work programmes. Without agreed definitions and quantitative measures of existing data, the value of data quality assessments may be lost. There is also the risk that valuable time and resources may be invested in completing lower priority data quality issues.

The existing Data Quality Assessment Methodology as a qualitative assessment of data elements needs to be revised in light of current information processes.

3. Proposal

We need to build on the foundation of data assets developed by the sector and provide a framework to proactively manage data quality. It is proposed that a Data Quality Framework is established to link the identification and prioritisation processes, assessment methodologies, recommendations, implementations, and reviews.

The purpose of the framework is to provide a common, objective approach to assessing data quality of all justice sector information. The framework enables the identification and measurement of data quality issues, standardises information, identifies priorities, and reviews data quality initiatives. This leads to the continuous improvement of data quality across the sector.

The Data Quality Framework allows for the consistent and accurate assessment of data quality. This will enable improved decision making and policy development across the justice sector. The framework will assist in the assessment of data quality at all levels of the justice sector using a data quality lifecycle.

A consistent assessment of quality over time will allow for the analysis of the effectiveness of data quality interventions, with assessment undertaken pre and post the intervention.

The first step in any improvement process must be to identify the uses made of the data and by whom. The framework also needs to look forward to the potential users of the data. The framework must consider the end user of the data and allow that user to define the level of quality required to make the data useful.

The Data Quality Framework should ensure that the customer:

- is able to access the data
- receives timely data
- finds the data are complete
- finds the data are accurate.

The Data Quality Framework will include robust qualitative and quantitative measures of data elements and provide an end-to-end business vision for data quality across the justice sector.

> *"Seemingly small data quality issues are, in reality, important indicators of broken business processes"* R. Kimball – An Architecture for Data Quality (2007).

The principles of data quality are concerned with the collection, storage, maintenance, and interoperability of data. Agencies have agreed the Data Quality Framework needs to apply to all data. That is, at all levels of the sector; business unit, agency, inter-agency, and justice sector irrespective of whether the data is shared or not.

The intention of the framework is to establish a sustainable, long-term quality improvement environment and culture both within sector agencies and across the broader sector. It is therefore necessary to move away from the current "one-off" assessment with recommendations approach.

The Data Quality Framework addresses the need for an end-to-end process for analysing and improving data quality across the sector. The framework also provides the adaptability to assess and respond to both urgent data quality issues and scheduled data quality assessments.

4. Data Quality Framework

Preparation of the Data Quality Framework has involved consultation with agency representatives on the Data Definitions Reference Group (DDWG) combined with a review of a number of justice sector, government, and data quality documents.

Sector representatives recognise the fact that data quality improvement is a substantial undertaking requiring ongoing effort. The framework is intended to provide a short and medium term solution by identifying specific initiatives to address priority issues. In the long term the framework provides for the ongoing maintenance and improvement of data quality.

To determine the components of the framework a number of existing quality improvement methodologies were researched. The foundation of the Data Quality Framework is based on the principles of a general improvement methodology, Six Sigma. This methodology has been chosen due to the focus on continuous improvement and supports the theory that data quality is a journey not a destination.

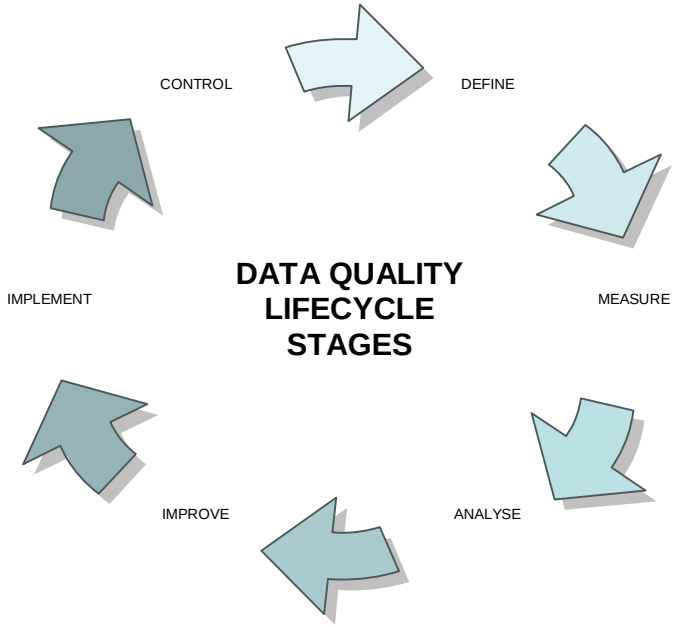
Six Sigma has two key methodologies: DMAIC (Define, Measure, Analyse, Improve, Control) and DMADV (Define, Measure, Analyse, Design, Verify). DMAIC is used to improve an existing business process and DMADV is used to create new process designs. In conducting this research, components of a framework appropriate for the sector were identified and resulted in the development of the Data Quality Lifecycle.

5. Data Quality Lifecycle

The Data Quality Lifecycle is the central component of the Data Quality Framework. Application guidelines, templates, and data quality resources (refer Appendix B) are supporting components of the framework.

The Data Quality Lifecycle for the justice sector consists of six stages; Define, Measure, Analyse, Improve, Implement, and Control. It should be noted that the Data Quality Assessment Methodology only partially meets the first four stages, *Define, Measure, Analyse,* and *Improve*.

Diagram 1 shows the Data Quality Lifecycle stages contributing to achieving data quality.



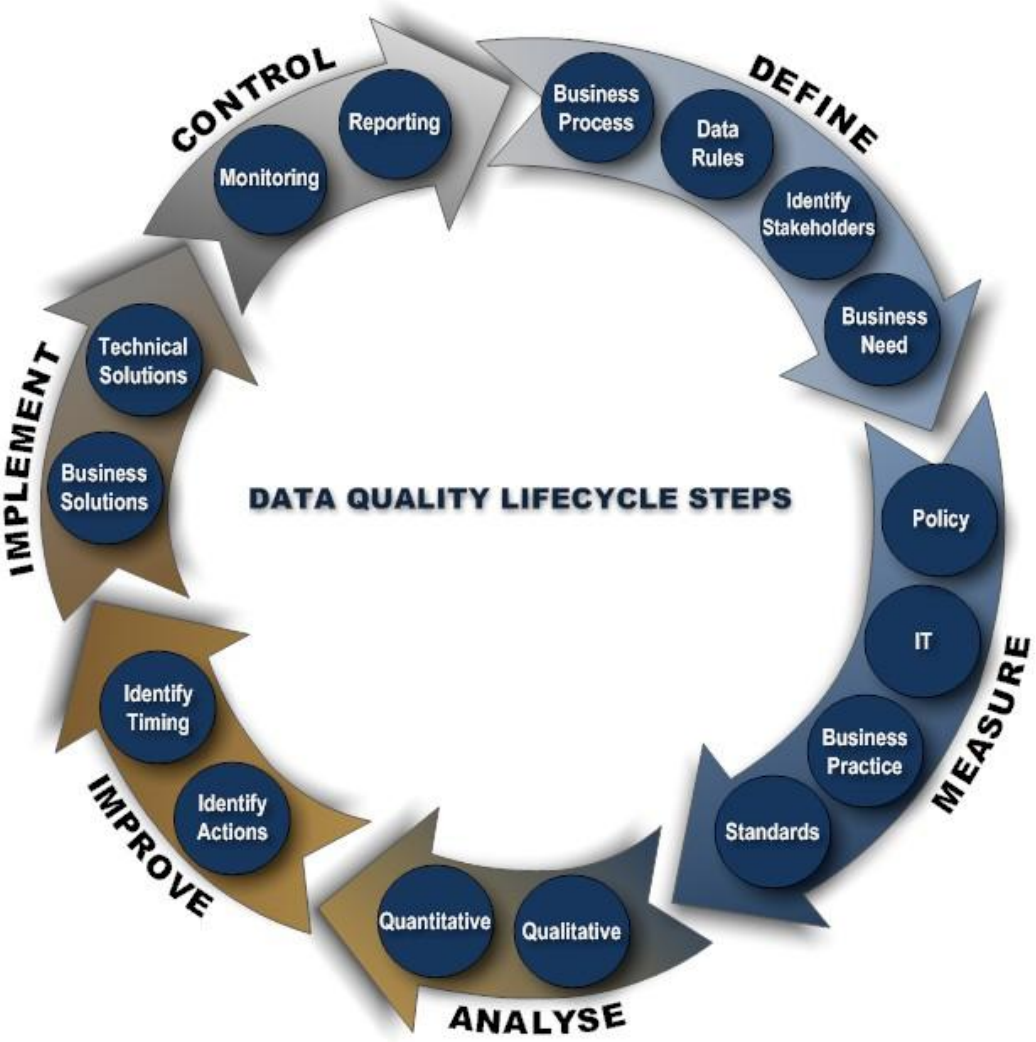
The high level applications of the Data Quality Lifecycle stages are described in the table below.

Data Quality Lifecycle Stage	**High Level Application**
Define	Define the business process, data rules, data goal, and identify the stakeholders involved with the data to be assessed. The errors and prioritisation, or business need, for data to be assessed will also be defined at this stage.
Measure	Measure the existing data in line with the data rules determined in the *Define* step. Areas to assess are policy, IT, and business practices as well as any standards or guidelines that the data should adhere to.
Analyse	Complete a gap analysis between the existing data and the data quality goal. Use both qualitative and quantitative measures as determined at the *Define* and *Measure* stages. Identify barriers to achieve the data quality goal and investigate areas of concern to the justice sector.
Improve	Design and develop an initiative to improve data quality based on the information determined at the *Analyse* stage. Identify the timeframe for implementing the initiative with consideration to team, agency, and sector work programmes and resource availability.
Implement	Implement the business and technical solutions determined at the *Improve* stage.
Control	Measure the data to assess whether the implementation of the initiative was consistent with the data rules and goal determined in the *Define* stage. The monitoring of data quality will be reported on a scorecard and will inform future data quality initiatives.

6. Data Quality Lifecycle Steps

Each of the Data Quality Lifecycle stages is broken down into a number of steps. The Data Quality Lifecycle steps are described in further detail in this section.

The diagram below provides a high-level overview of the Data Quality Lifecycle steps that agencies must consider when implementing initiatives impacting on data quality both within agencies and across the justice sector. To prompt and test data quality policies, a detailed Data Quality Checklist is provided in Appendix E.



6.1 Define Stage

There are four steps in the *Define* stage of the Data Quality Lifecycle. The four steps are *Business Process, Data Rules, Identify Stakeholders,* and *Business Need*. The objectives of this stage are to define the business processes, identify stakeholders and determine data rules and data quality goals. The prioritisation of data to be assessed based on business needs and known issues and/or errors, will also be defined at this stage.

6.1.1 Business Process

This step requires the understanding and modelling of the flow of existing data. This includes data collection, maintenance, and storage. This step should consider all of the intended uses and interoperability of the data.

The key issues to consider are:

- What purpose is the data collected for?
- What business processes are in place to control data collection, usage, and change?

Example
Police provide their communications centre with ethnicity to assist in the identification of suspects and wanted persons, etc. Ethnicity data is also captured at the time of arrest and entered on both a paper and electronic form. This data is used for policy and strategic decision making and is shared with Justice and Corrections. Any changes to the Statistical Standard for Ethnicity 2005 are notified through the Justice Sector Change Control Notification (JSCCN) system to assist the sector's compliance.

6.1.2 Data Rules

The *Data Rules* step involves the definition of terminology, data attributes, and counting rules. For example, any data that meets the data definition and does not meet the data rule is counted as an error. There is a need to determine if the resulting data quality issue relates to a single data element (e.g. ethnicity) or a collection of data elements (e.g. identity).

The key issues to consider are:

- Have data rules for the data been defined and documented?
- Can data be broken down into smaller components, or can it be aggregated?
- Have agreed data definitions been used? If not, why not?
- What is the data quality issue?
- Is the data quality goal in line with the data rules?

Example
The ethnicity code is defined in the sector Data Dictionary as "Identifies the ethnic or social group a person most associates with. Where possible this should be identified by the individual, although in the justice sector this may not be possible." An acceptable ethnicity recorded on the charge sheet will be the corresponding five digit code value. The sector has an agreed data definition and data rules for ethnicity. The data quality issue is that a number of text fields and null values are being shared across the sector.

Sentence and parole conditions are examples where text fields are sent across interfaces. Often these tend to run together and hence are potentially difficult to interpret. The field lengths are also limiting.

6.1.3 Identify Stakeholders

To ensure the business processes are fully understood, representatives from all areas of the business, whether agency or sector focused, should be identified. Involvement of these representatives in the data quality improvement process will strengthen data integrity within the agency and sector.

It is important to remember that data creators and data customers/consumers work in different parts of each agency. Problem solving requires an appreciation for the 'horizontal' processes. That is, to achieve data quality, all areas of the agency and or sector, must be consulted with and involved in the definition and resolution of data quality issues.

The key issues to consider are:

- Where is the data sourced from?
- What other departments or agencies use existing data?
- Will new data collections be used by others - now and/or in the future?
- Have stakeholders in different parts of the agency/sector been consulted with e.g. Policy, IT, and Operational?
- What are the intended uses of the data? (end-to-end process)
- What is the level of data quality required? (determined by stakeholders)

Example
Stakeholders for criminal justice data include, police officers tracking particular offenders and offences, and justice sector management interested in measuring the level of crime and how it is dealt with, and statisticians (Police, Justice, Corrections, and Statistics NZ).

Stakeholders for conviction and sentencing data include court officers tracking cases, senior executives interested in court workloads and outcomes, and statisticians (Police, Justice, Corrections, and Statistics NZ).

6.1.4 Business Need

This step analyses the business need/priority for the assessment to take place. It is accepted that there are some data quality issues which require urgent or immediate attention and some which are prioritised and scheduled. It may be necessary to consider the impact to the business when determining the priority of data quality assessments to be completed (scheduled) and apply a weighting to the different areas of the business.

The key issues to consider are:

- What level of data quality is required for management to feel confident in strategic decision making?
- Is the data shared between agencies?

Example
Scheduled – a scheduled data quality assessment of offence codes was completed in October 2004. The assessment of offence codes was chosen due to the fundamental role that offence codes have across the justice sector.

Urgent – the removal of data identifying individuals from a field that is shared and accessible.

6.2 Measure Stage

There are four steps to the *Measure* stage of the Data Quality Lifecycle. The four steps are *Policy, IT, Business Practice, and Standards.* The objective of this stage is to measure existing data in line with the data rules determined in the *Define* step. This stage involves measuring what currently exists, including key aspects of the current process and collecting relevant data in the areas of policy, IT, and business practices. Where appropriate, standards and guidelines for the data to adhere to, should also be measured.

6.2.1 Policy

The objective of the *Policy* step is to understand and document government, justice sector, and agency policy relating to data quality, uses of data, and governance of data.

The key issues to consider are:

- Have all stakeholders' policy needs been considered?
- Is the existing policy up to date and in current use?

Example
Sometimes errors are discovered when the administrative data are converted into data for statistics. These are often corrected within the statistics, but the business should have a policy on whether they are corrected at the source.

6.2.2 IT

The *IT* step involves the use of technical tools to determine the current quality of data. The use of profiling software or sampling of data quality will result in a quantitative measure to determine if a data quality issue exists and to what extent.

The key issues to consider are:

- Is the data structured or free text?
- Are there null values?
- Is the data complete?
- Is the data sourced from a pre-populated field?

Example
A step in the process of converting administrative data into data used for statistics should be to produce one-way frequencies of the major variables. These frequencies can be checked for inaccurate or new codes (e.g. new offence codes, or negative ages because a date is incorrect). Two-way frequencies may be used to check inappropriate combinations (e.g. a life sentence for using cannabis). These problems should be corrected on the statistics collections, but it is sometimes difficult to get corrections on the administrative data source.

6.2.3 Business Practice

For the *Business Practice* step it is important that an end-to-end data flow process is measured. This is all of the business practices involved in the data collection, maintenance, storage, and if applicable, the interoperability of the data. This step does not solely focus on the technical specifications required for IT systems.

The key issues to consider are:

- Is the data shared?
- Have sources of the data been verified and listed?
- What is the data used for and is this fully documented?
- If used in reports, which reports and how frequent?

Example
Personnel who work in the justice sector over a period of time develop an in-depth knowledge and understanding of sector processes. Often this is not documented and when staff leave the organisation/s that institutional knowledge is lost.

6.2.4 Standards

The purpose of this step is to ascertain if there are any external standards to be complied with or considered. In doing this, upcoming changes to existing data will be considered, resulting in a reduced need for assessment and re-work.

The key issues to consider are:

- Are there any external standards required to be met?
- Will the data be used by others – now and in the future?

Example
Statistical Standard for Ethnicity 2005.
Crime data, conviction and sentencing data are Tier 1 Official Statistics (a defined set of key official statistics that are performance measures of New Zealand). As such they are required to use standard classifications (e.g. for gender, and age) in any statistical reports so that they are comparable with other Tier 1 Official Statistics.

6.3 Analyse Stage

There are two steps to the *Analyse* stage of the Data Quality Lifecycle. These are *Qualitative* and *Quantitative*. The objective of this stage is to measure the gap between the data quality goal defined at the *Define* stage and the results of the *Measure* stage and identify barriers to achieving the goal. A gap analysis using qualitative and/or a quantitative analytical tools is necessary. The appropriate type of analysis to be completed will have been determined at the *Define* stage.

6.3.1 Qualitative

Qualitative measures may include the use of interviews and/or the Data Quality Assessment Methodology. Agencies should investigate all areas of concern highlighted through the *Define and Measure* stages. Barriers to achieving improved data quality should also be identified through this process.

The key issues to consider are:

- Has the information been created or collected from an accurate and relevant source?
- What processes are in place to control data collection, usage, and change?
- Has a new computer system been developed over the lifecycle of the data?
- Have forms for data entry changed?

Example
The decommissioning of the Law Enforcement System over a number of years must be factored into any data analysis which incorporates data between 1998 and 2005.

6.3.2 Quantitative

Quantitative analysis is the number-crunching between the quantifiable data quality goal determined at the *Define* stage, and the current position, determined at the *Measure* stage. Using both the *Business Practice* determined at the *Measure* stage and quantitative analysis, the source and extent of any data quality issues identified may become apparent.

Tools such as a scorecard will assist individual agencies and the sector to understand the impacts of data quality (refer 6.6.1).

The key issues to consider are:

- What is the quantifiable measure of existing data quality?
- How has the data quality changed over time?

Example
In a data quality assessment some variables have a high percentage of missing values and are therefore not used for statistics. For example, in conviction and sentencing data produced from LES the "plea" variable had a high percentage of missing values, and also a high percentage of the value "other" – which should have been used rarely.

This has now been rectified with plea being a mandatory field.

6.4 Improve Stage

There are two steps in the *Improve* stage of the Data Quality Lifecycle; *Identify Actions* and *Identify Timing.* The objective of the *Improve* stage is to optimise data quality based upon the outcome of the *Analyse* stage. This involves making recommendations to improve data quality including coordinating work programmes and establishing the timeframe for implementation.

6.4.1 Identify Actions

The opportunities/actions identified to improve the data quality issue will be based on the outcome of the *Measure* and *Analyse* stages of the data quality lifecycle and the prioritisation given to the data element(s). Understanding where the data quality gaps and issues are will inform the types of recommendations made for improving data quality. The key requirement of this stage is to ensure thorough consultation across the business, including operational, technical, and end users of the data element.

The key issues to consider are:

- Have operational, technical, and end-users been involved in identifying actions to improve data quality?
- Does the data need to be stored in a structured way?
- Will changes enable operational or analytical needs to be met?

Example
Replace free flow text field with structured and defined fields that meet the needs of all end users and stakeholders. This is however not always possible.

Use field and screen validations to support the business process.

6.4.2 Identify Timing

It is important to gain small wins and maximise opportunities to leverage resources within an agency or across the sector. All opportunities to consult, understand, and plan changes in a timely fashion must be considered to achieve maximum buy-in of the data quality initiative.

The key issues to consider are:

- Scheduled work that will impact the change
- Resource commitment across the business from all agencies
- Work programme items that changes can be attached to
- Legislative changes
- Software delivery lifecycles
- Change management policy and training requirements.

Example
Problems with the “plea” variable were fixed when CMS replaced LES.

6.5 Implement Stage

The objective of the *Implement* stage is to implement the tasks outlined in the *Improve* stage. There are two steps in the *Implement* stage *Business Solutions* and *Technical Solutions.* Improvement initiatives may include both business and technical solutions. The solution to data quality issues must always consider the needs of the business and not have a sole technical focus.

6.5.1 Business Solutions

The intention of the *Business Solutions* step is to implement the actions identified to meet the needs of the business. These solutions are not technical. Solutions may include training, business process improvement, documentation and escalation steps identified.

The key issues to consider are:

- How will changes be implemented and communicated?
- Is there an accountability structure for the data process?

Example
Update documentation within JSCCN process and policy manuals

Implementation of the Police National Recording Standard for police data collection.

6.5.2 Technical Solutions

Technical solutions must be implemented in consultation with the business users identified in the *Define* stage. This document does not override any agency change control policy and therefore must consider any change management necessary at both the agency and sector level.

The key issues to consider are:

- Has documentation been completed, authorised, and stored?
- Has an audit trail, for changes made, been incorporated into the systems design?

Example
Uploading the amended notes and controlled value lists (CVLs) into the JSCCN system.

Uploading the amended policy and process manuals onto the JSCCN system.

6.6 Control Stage

There are two steps in the *Control* stage of the Data Quality Lifecycle; *Monitoring* and *Reporting*. The objective of this stage is to monitor the continuous improvement of data quality and to identify any decline in data quality to inform future assessments.

6.6.1 Monitoring

The purpose of the *Monitoring* step is to determine if the actions implemented have resulted in the intended consequences. The intended consequences are identified in the *Improve* stage. If the data quality goal is quantifiable then a scorecard approach can be effective. A scorecard can be used to monitor the before, after, and ongoing status of data quality.

In some instances data quality may be adversely affected due to unintended consequences of other changes made. It is the intention of the framework to limit and reduce the likelihood and occurrence of unintended consequences on data quality.

The key issues to consider are:

- Are the definition and data rules determined at the Define stage still relevant?
- Did the implementation deliver the data quality goal defined?
- Have lessons learnt been documented for future data quality assessments?
- In ongoing monitoring, are the same data quality issues occurring or are new issues arising?

Example
Statistics should include the number of unknown values for a variable.

For example, in 2006 there were 1,570 convicted cases where the age was unknown (1.4% of all convicted cases in 2006).

6.6.2 Reporting

On completion of the *Monitoring* step the measurement should be added to the data scorecard for review on a time period determined at the *Analyse* stage. It may not be feasible to achieve the data quality goal determined by the end user at the *Define* stage on the first iteration of the data quality lifecycle. In these instances it is recommended to set a more realistic short term goal and gain smaller wins over a longer period of time.

The results of the Monitoring step must be reported to the stakeholders involved in the *Define, Measure,* and *Improve* stages of the lifecycle. By maintaining visibility of data quality on a regular basis, opportunities for continuous improvement in data quality will be realised and enthusiasm for data quality maintained.

A scorecard can reflect differing views i.e. individual and combined data elements with a proposed target, which will have been determined in the *Define* stage of the data quality lifecycle. The figures that would be recorded are from the *Measure* stage. The impact of data quality initiatives will be captured through the ongoing *Control* stage of the data quality lifecycle. Data Element and Data Quality Improvement Scorecards are provided as examples in Appendix G.

Key issues to consider are:

- What is the appropriate reporting mechanism?
- What frequency is reporting required?
- What are the reporting needs of the audience?

Example
Scorecard – As no quantitative analysis has been included in data quality assessments within the justice sector, actual data is not available. Regardless of how good or bad the current position may be, a "stake in the ground" view needs to be set, so that monitoring can commence.

7. Application of the Data Quality Framework

The Data Quality Framework applies to the definition, assessment, improvement, and monitoring of data quality across the justice sector. It is designed to identify and assist resolution of any data quality issues determined. Data quality issues may be entirely within an agency or at a sector level. Irrespective of the size of the data quality issue the Data Quality Framework will guide the improvement of data quality.

There are two primary types of data that the Data Quality Framework can be applied to; *New Data* and *Existing Data*. The differing processes that may apply are mapped in the Data Quality Assessment Flowchart (Appendix F).

7.1 New Data

When there is a requirement for new data to be collected, all six stages and sixteen steps of the Data Quality Lifecycle will need to be completed. It is expected that the requirements of the Data Quality Lifecycle would be incorporated within individual agency system development lifecycles. This can be achieved by incorporating the lifecycle as is, or an agency's lifecycle that covers all these elements (ie an agency may refer to the step by a different name, provided it encapsulates the requirements of the framework).

7.2 Existing Data

There are two applications of the Data Quality Framework for existing data, scheduled and urgent. The application is dependent on the priority the business places on the data quality issue.

7.2.1 Scheduled

For a scheduled data quality assessment the full data quality lifecycle must be worked through. This is to ensure that all the data quality lifecycle stages and steps are covered effectively, ensuring that any data quality issue will be identified. It is important that the *Define* and *Measure* stages, including data quality definitions and business rules, are up to date in the current environment.

Application of the complete Data Quality Framework will enable the associated resource commitments for sector staff, to reduce over time. Benefits gained from a structured approach, agreed business rules, appropriate controls and knowledge documented, will reduce the effort required. At the same time, this will improve data quality.

7.2.2 Urgent

In the event of an urgent data quality issue, an immediate assessment and solution is required. There is generally, no time to identify all stakeholders as a remedy is required within a matter of hours or days.

When an urgent data quality issue has been identified the following actions have been identified as the minimum requirements. The actions (and associated steps from the Data Quality Lifecycle shown in italics) to be taken are:

Minimum Requirements	**Data Quality Lifecycle Steps**
Identify the extent of the problem	*Define, Measure, and Analyse*
Identify the risk to the business	*Define, Measure, and Analyse*
Determine the cause of the issue	*Define, Measure, and Analyse*
Determine a proposed solution	*Analyse and Improve*
Escalate the problem to a senior manager with a proposed solution	*Improve*
Coordinate with the manager to contact senior stakeholders	*Define and Improve*
Implement the agreed proposed solution	*Implement*
Document the issue, the steps implemented, and refer to the Data Definitions Reference Group	*Control*

Appendix A. Data Quality Dimensions

From previous work completed on data quality in the justice sector, fifteen data quality dimensions were identified to assess the quality of data, dependent on the context and use of the data. The fifteen dimensions are: Accuracy, Believability, Objectivity, Reputation, Completeness, Timeliness, Appropriate amount, Relevance, Conciseness, Consistency, Interpretability, Understandability, Accessibility, Ease of operation, and Security.

Of the dimensions identified, the data quality dimensions of primary interest to the justice sector are outlined below. The first four data quality dimensions (highlighted in bold) have been identified as being of the highest priority in the assessment of data quality.

Term	Definition
Accuracy	The quality of the content equates to the degree to which the data values accurately describe the meaning of the real-world fact and how well values conform to business rules.
Completeness	Data is complete across all the database records for which it is required.
Timeliness	Data is available when it is required. Includes the acquisition, delivery, and use of data.
Accessibility	Information can be easily found and accessed by those with a requirement to use it. Potential users of data are aware that it exists, and data security is set at an appropriate level.
Representation	Data is provided in a format that is useful to those who require it.
Definition	The meaning or specification of the data and the associated business rules.
Consistency	Data is consistent in definition and treatment both within and across databases.
Relevancy	Data is used to accomplish the work and mission of the agency. The type of data specified for collection is useful, without missing data elements.

The data quality dimensions used to measure data quality will be dependent on the type of data under consideration.

Appendix B. The Data Quality Toolkit

The Data Definitions Reference Group (DDWG) have established resources to ensure that shared information is collected, stored, maintained, and interoperable in a standardised format.

This toolset of resources contribute significantly to building a strong foundation for achieving high quality data and/or information across the justice sector. The Data Quality Toolkit currently consists of the following:

- The Justice Sector Data Dictionary – 1998 (living document)
- Justice Sector Change Control Notification (JSCCN) System – October 2002
- Data Quality Assessment Methodology – September 2006
- Justice Sector Information Stocktake: What's Where – March 2007
- Data Management Catalogue: For the New Zealand Justice Sector – May 2007
- Privacy Act Training Resources – May 2007
- Information and Knowledge Management Reference Kit – November 2007
- JSCCN Policy Manual – January 2008
- JSCCN Process Manual – January 2008.

Appendix C. Poor Data Quality Causes

Poor data and information quality can create chaos. Unless the root cause is diagnosed, efforts to address it are akin to patching potholes. On first inspection the causes of poor data and information may be readily apparent. However, it is important to get behind the immediate causes to discover the root causes of the problem. A common misconception is that data quality problems are caused by data producers. Whilst at a superficial level it may appear that fault lies with the operational staff entering data, further analysis may uncover underlying issues such as training, incentives, and system limitations. Addressing causes of poor data quality requires a co-operative effort. International data quality expert Larry English notes:

> *"Information quality improvement requires a blame-free and non-judgemental environment. 'Fault finding' only creates fear and stymies creative change. It leads people to cover up 'problems' and not to be open to exploring process improvements"*

High Level Causal Factors

Lack of business incentive to produce quality data.
This is often the case where the agency responsible for generating the data is not the end user of the data, leading to limited motivation and lack of understanding/knowledge of the use of the data. The requirement to produce quality data can also conflict with operational tasks and the basic mission of the agency responsible for collecting the data.

Lack of individual incentive to produce quality data.
For example, include customer service representatives who are rewarded for the number of customers handled with no incentive given to the collection of accurate data.

Lack or limited conformance to standards accompanied by a lack of documentation.
Because data is primarily collected in support of operational and short lived activities, little attention is paid to developing standards or maintaining documentation within the data capturing business units.

Multiple sources of the same information.
Producing the same information using several different processes is likely to produce different values for the "same" information.

Lack of information available to the information users regarding how the data was collected.
Information stored in organisation databases is often considered as factual. The collection of these facts may, however, involve subjective judgement. The risk is that a lack of understanding of the information may lead to incorrect decisions.

Inadequate training
Inadequate training combined with system deficiencies, such as missing field validation and loose business rules, can lead to systemic errors in information production.

Poor identification and analysis of information users
A poor understanding of the needs of data and information users may impact what data is collected, how it is defined, and stored. Even if data is of high quality in most areas,

if potential users know nothing of its existence or are unable to gain access, its value is severely limited.

Poor accessibility leading to a lack of data use beyond case processing
Poor access to data beyond front-line operational use leads to reduced identification of records that are inconsistent, outside policy or missing key values.

Information Quality Matrix
The following table provides a starting point for analysing data quality. The table identifies possible causes of, and possible resolutions for common data quality problems.

Prioritise data quality issues by identifying issues with multiple contributing possible causes and then weighting these causes to produce a score – highest score equals the number one priority for data quality assessment.

Problem	Possible Causes	Possible resolutions
Definition The definition of the data does not describe the meaning of the real-world fact	• inadequate analysis of the data users and their requirements for the data • inconsistent definitions across unit/agency/sector boundaries • poor documentation • definitions not made available to the information users.	• review understanding of the data users and their requirements of the data • implement agency and sector standard definitions • where appropriate use industry or all-of-government standards • document definitions and make readily available.
Accuracy The data value does not describe the real-world fact	• poor conformance to business rules • poorly defined business rules • operational definitions not clear to data users/information producers • operational staff measures and incentives give little importance to quality data • changes in real-world records not updated • new record created rather than existing record updated • incorrect value provided by end "customer" • data entry error - mistype • misuse of data fields due to work-arounds • data corruption.	• modify system to enforce business rules • limit the use of free text fields • improve user interface design • provide appropriate level of system help • monitor data quality • include data quality in performance measures and position descriptions • review relevancy of data collected at source • improve searching and customer matching capabilities • improve visibility and access of data to customer • improve data quality training • instigate data cleansing activity to correct existing problems whilst ensuring the causes are adequately addressed.
Completeness Data required is not complete across all database records for which it is required	• operational constraints restrict ability to collect data • data not available at time of transaction • inadequate use of mandatory fields • lack of historical record retention • inadequate training.	• review data collection process • where appropriate use statistical analysis to provide an explicit level of certainty around information products.

Problem	Possible Causes	Possible resolutions
Consistency Data is inconsistent in definition or treatment within and across databases and agencies	• poor quality database design • lacking data definition and architecture standards • data collected more than once and maintained in more than one place • historical separation and merging of business units or agencies.	• implement agency or sector standard definitions • consolidate data entry and maintenance functionality • document uses.
Relevancy The data is not used or useful to accomplish the work and mission of the agency	• inadequate analysis of the information users and their requirements of the data • analysis not updated to reflect changes in agency requirements over time.	• review understanding of the information users and their requirements of the data • archive data no longer required.
Accessibility The data is not readily available to the information users who require it	• privacy and security policies restrict access • lack of knowledge of what data is available and or how to obtain it • lack of system access.	• review policy and implementation to ensure correctly applied • provide access to information through a widely accessible medium (e.g. intranet, Internet).
Timeliness The data is not available within the timeframe it is required	• inadequate analysis of the information users and their requirements • batch processing scheduling is too slow.	• re-prioritise the processing and provision of data.
Representation The data provided is difficult for the information workers to understand and use	• inadequate analysis of all the information users and their requirements of the data • the description of information and the data from which it is built is inadequate.	• review "form and format" requirements with information users • automate processing of data into usable presentation formats.

Appendix D. Poor Data Quality Costs

There are two areas to consider when determining the cost of quality to an agency; the costs resulting from failure to assure quality, and the costs to assess and improve the quality of the process and resulting products.

Data Quality Assurance Failure Costs

These are avoidable costs due to process failure, information scrap and rework, and lost and missed opportunities. These include costs due to:

Process failure:

- errors in operational decisions
- frequent system failures and service interruptions
- protection from and exposure to liability and compensation
- loss of internal confidence in the information produced
- management costs (characterised by reactive measures) associated with addressing loss of external confidence (users/public) in the service provided
- lack of confidence contributes to increased frustration within the agency, low moral, and increased staff turn-over.

Information scrap and rework:

- undoing work that has already been done
- data re-collection and correction including the cost of data cleansing activities
- re-execution of the data dependant activity
- workaround costs and decreased productivity - including costs of double checking the information caused by the loss of internal confidence in its quality.

Lost and missed opportunities:

- poor policy decisions
- missed outcomes
- missed funding opportunities.

Quality Assurance and Process Improvement Costs

These are costs to assure processes are performing properly and to improve processes to prevent defects arising. They include costs due to:

Inspection and assessment:

- data quality inspection and analysis software and hardware
- data quality inspection and analysis people.

Process improvement and defect prevention:

- process improvement initiatives
- education and communication.

Often the true costs of data quality are difficult to establish, with the impacts of poor quality information leading to “hard” dollar or “soft” intangible costs. At first glance doing something about the status quo can appear an expensive exercise which might include quality analysis, data cleansing projects, and ongoing initiatives.

The case to establish data quality improvement initiatives and justify expenditure is best made by highlighting the high costs of doing nothing.

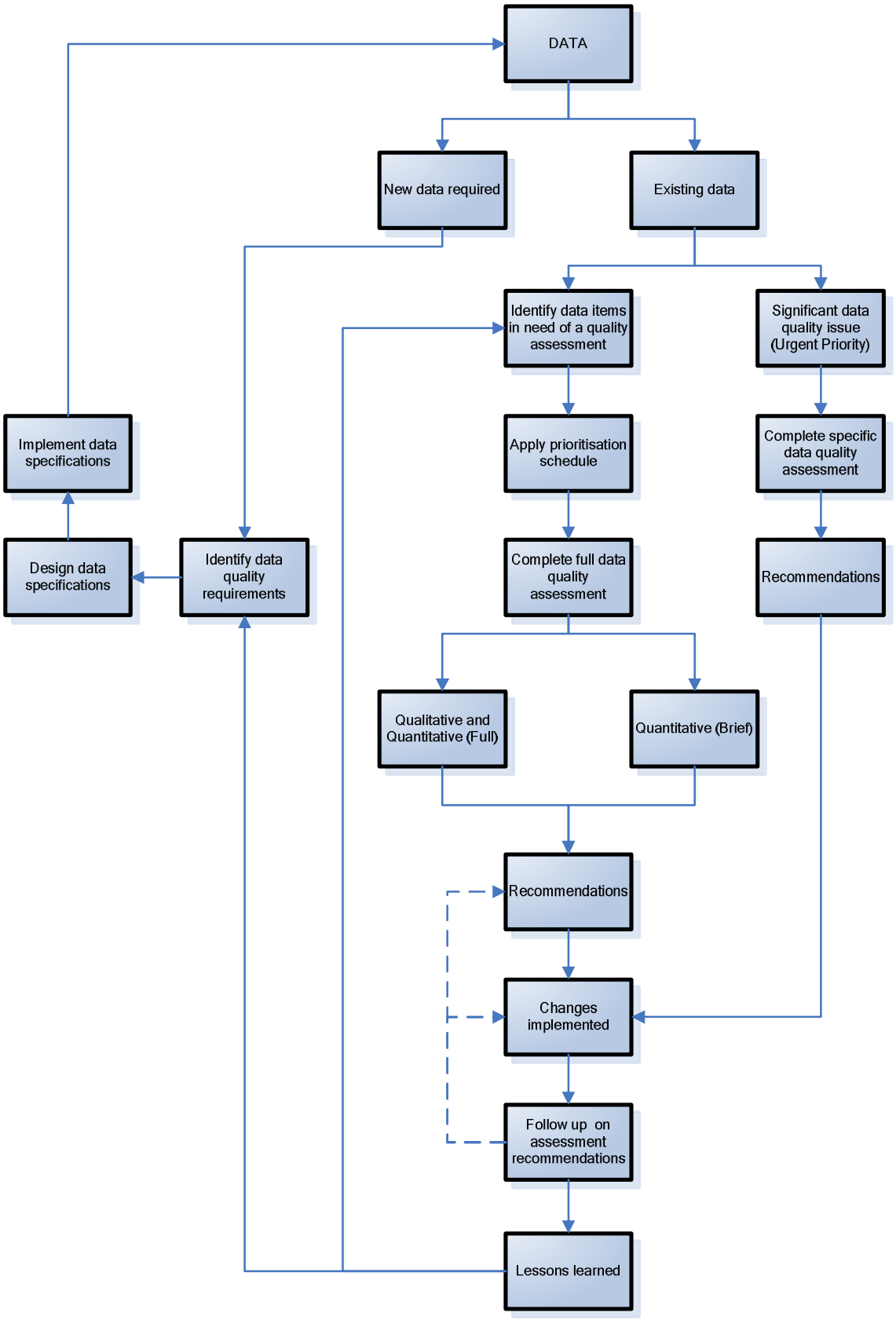
Appendix E. Data Quality Checklist

This is a checklist for agencies to use when designing new data for collection, amending existing data, or in assessing data quality. The following checklist is intended to prompt and test data quality policies and procedures for completeness, robustness, and compliance with current standards and good practice.

Check	Checklist
	Accuracy
	Duplicate Records
	Is there a process to check for and remove duplicate records?
	Validation Checks
	Are addresses verified during input?
	Are addresses geo-coded?
	How are inconsistencies resolved?
	Is data quality monitored and how often?
	Integrated Data
	Is data linked between systems or agencies?
	If so, is there a verification process that runs on exchanged data?
	Is data benchmarked against any other source?
	Consistency
	Is the data consistent?
	Null values
	Are there null values?
	How are null values to be handled?
	Aggregations
	Is any data aggregated or rounded? If so, is it documented?
	Missing Records
	How are missing records dealt with?
	If data is missing, is it flagged with a certain value?
	Classifications and Metadata
	Classifications
	What classifications are used for the following: Age, Gender, and Ethnicity
	How is the data split into regions?
	Does the data need to be stored in a structured way?
	Metadata
	Is any field-level metadata available?
	Is there any process documentation available?
	Collection
	Data Entry
	How is the data collected?
	Is the data derived?
	Are there gaps in the collection process?
	Has policy changed the type of data collected? If so, how?
	Is data entry a consistent process?
	Has the information been created or collected from an accurate and relevant source?
	Collection Frequency and Period
	How often is data collected?
	How long does collection take?
	Is there a time interval between data collection and data input?

Check	Checklist
	Usage
	What business processes are in place to control data collection, usage, and change?
	What data is sourced from other agencies?
	Have sources of the data been verified and listed?
	What other agencies use the data?
	Who is the audience that the data is intended for?
	What data do I have relevant to my needs?
	What data do others have relevant to my needs?
	What data do I have relevant to other sector agencies needs?
	Will the data be used by others – now and future?
	General Information
	Who is the custodian of the data?
	For what purpose is the data collected?
	What is the business need for the data?
	Is there a clearly defined outcome for the agency/sector that this data will support?
	Is the purpose of the data clear? E.g. Is it possible for the data to be used out of context?
	Have all relevant stakeholders been identified and involved? E.g. Policy, IT, & Operational?
	Is the data shared between agencies? If so, use the JSCCN system to notify change
	Is there an accountability structure for the data process?
	Data Rules
	Have data rules been defined and documented?
	Have standard/agreed data definitions been used?
	How will data quality be measured?
	What is the acceptable level of data quality for use?
	What is an exception or error?
	How are exceptions or errors to be handled in reporting?
	Are there external standards to be adhered to?
	Time Series
	Would it be possible to carry out a time series analysis of the data? If so, how far back?
	Does the system force selection of options? E.g. Reason field, blank not accepted
	Is there an ability to add attributes easily?
	Change Management
	Is there a change management process for the data?
	Can users raise issues for change management?
	Have forms for data entry changed?
	Has a new computer system been developed over the lifecycle of the data?

Appendix F. Data Quality Assessment Flowchart



Appendix G. Data Quality Scorecards

The tables below are examples of how a data quality can be monitored and reported.

The Data Element Scorecard would record the result of assessment of single data elements from the Measure stage of the Data Quality Lifecycle.

The Scorecard of Data Quality Improvement would record the assessment of combined data elements. For example 'identity' is generally made up of a person's name, date of birth, and sex. Each of these on its own is a data element. It is not a true measure to try and understand a collection of elements without first breaking down the combined group to the simplest form (data elements). Without understanding 'identity' the quality of the data can not be improved from a position of knowledge.

Data Element Scorecard - *example*

Data Element	DQA Date	2006	2007	2008
Ethnicity	Sep 2004	56%	58%	75%
Offence Code	Oct 2004	76%	72%	87%
Date of Birth	Sep 2008	58%	62%	64%
Sex	Mar 2009	72%	89%	89%

Scorecard of Data Quality Improvement - *example*

Data Item	DQA Date	2008 - current	2009	2010
Ethnicity	Sep 2004	75%	85%	Target 95%
Offence Codes	Oct 2004	87%		Target 95%
Safety Alerts	Feb 2005	82%		Target 98%
Identity	Apr 2006	63%		Target 85%