Data Quality Control Practices for Implementation in the Queensland Water Industry

Produced by ***qldwater***

May 2011

Address: Level 1, Local Government House,
25 Evelyn Street, Newstead Qld 4006
Post: PO Box 2100, Fortitude Valley BC Qld 4006
Phone: +61 7 3252 4701 **Fax:** +61 7 3257 2392
Email: enquiry@qldwater.com.au **Web:** www.qldwater.com.au

Table of Contents

Introduction 3
What is Quality Control (QC)? 3
Why Quality Control? 4
The Development of this document 4
Survey 4
Workshop 4
Overview: How to use this document 5
The QC Template 5
A Note for SWIM Users 5
How to Check Data 5
What it could mean if your data is failing QC Checks 6
QC Practices: Descriptions, Instructions and Examples 6
1. Comparison to historical data 6
2. Comparison to other schemes and SPs 7
3. Relational Data Checks 8
4. Derived Values 9
5. Numeric/text format 10
6. Acceptable Ranges 11
7. Ineligible Zero Values 11
8. Validation 12
9. QC checks in sampler/device 13
10. "Eyeballing" 13
SP Practices to Increase the Value of QC 14
Institutionalised Quality Control Protocols 14
Data Management & QC Training 14
Data Dictionary/Thesaurus 15
Users forum to exchange data knowledge 15
Different Data Categories and the Relevant QC Checks 15
Table 1: Grouped by data category 16
Table 2: Grouped by QC Checks 17
Acknowledgments 17

Introduction

The Queensland Water Directorate (*qldwater*) is the peak body representing Service Providers (SPs) throughout the Queensland water industry. In addition to advocating on behalf of SPs, *qldwater* seeks to provide its members with facilities to directly aid them in providing high level services in the urban water industry.

The intention of this work is to describe the best-appropriate practices for QC, to increase the efficiency, timeliness and accuracy of reporting by Queensland's SPs. Rather than identifying 'best practices', which may not be the 'best' for the situation of every SP or dataset, *qldwater* has set about identifying the best-appropriate practices. To ensure the practices identified are appropriate, the water industry has been involved in the identification of the QC practices.

In addition to identifying the 'best-appropriate' practices, *qldwater* has drawn from practices currently in use by industry. These have been supplemented by additional practices developed in-house by *qldwater.*

This document ("Data Quality Control Practices for Implementation in the Queensland Water Industry") describes a set of practices that has been identified by *qldwater* in consultation with SPs. This document aims to provide guidelines describing the use of those quality control practices, in a form that users may immediately adopt to improve local Quality Control Practices. This information is being provided in the form of a document that is ready to be incorporated in to SP operational procedures to make it easier and faster to implement QC.

This quality control project was supported by funding from the Bureau of Meteorology (BoM).

What is Quality Control (QC)?

In reference to the data that is collected in the Queensland Urban Water Industry, the term 'quality control' will be used to describe systems put in place to ensure that data which has already been gathered is as accurate as possible. Quality Control means the practice of checking data. This differs from processes which could be put in place to ensure data is accurately generated in the first instance, such as regular calibration of meters, which seek to reduce the number of inaccuracies in data as it is produced. While these data calibration processes are vitally important, they are not within the scope of this document. Quality Control will be used to describe processes that aim to eliminate errors or mistakes in data that has already been produced.

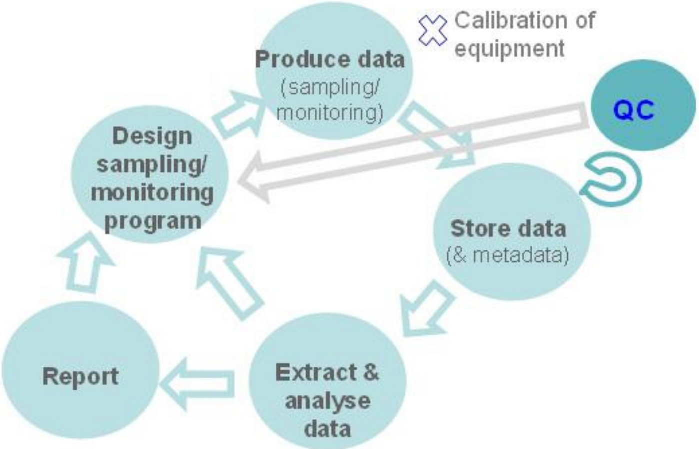

Figure 1: Data Cycle

Figure 1 shows a general "data cycle"; a flow chart describing the path of data from collection to reporting/use. QC, as indicated by the circular arrow, is conducted on stored data. The direct results of data quality control are applied to this same stored data, i.e. if a mistake is found, it is immediately corrected. In the long-term, quality control practices may reveal a pattern of errors, which would indicate that the sampling/monitoring program design may need to be revisited (as indicated by the grey arrow in Figure 1).

Why Quality Control?

For data to serve any purpose at all, it must be reliable. Any procedures that can increase the accuracy of data improve its overall reliability. This has benefits for all users of the data; the SPs and their governing councils, industry regulators (e.g. Department of Environment and Resource Management) and any other parties involved in the broader water industry (e.g. the Bureau of Meteorology, the Australian Bureau of Statistics). The ultimate benefits from improving the quality of water data are better management and operation of SPs, better informed decision-making by government bodies, and better water services for the people of Queensland.

Access to a defined QC procedure will also create efficiencies in that incorrect data can be identified before it is used for analyses and reporting; when incorrect data is discovered at a later time, extra time is required to reanalyse and correct reported data.

The Development of this document

Survey

Before making a recommendation about QC practices for SPs, ***qldwater*** conducted a survey to establish which practices were already in use. Representatives from 14 SPs spoke with ***qldwater*** via telephone to provide confidential information about current QC practices. Thirteen of the fourteen SPs questioned identified that some level of quality control was applied to data in their organisations. All SPs indicated that there were significant quantities of data that were not subjected to quality control and all expressed a strong desire for improved data accuracy.

Workshop

Using information gathered during the survey, as well as independent research and experience gained through the ongoing SWIM program, ***qldwater*** developed a list of data QC practices that was presented to attendees of a workshop on 30 March 2011. This workshop was attended by representatives from over 20 Water Service Providers (LG-owned/operated), DERM (Water service regulator/data users), BoM (Data users), as well as presenters; staff from ***qldwater*** and independent data quality experts consulting to ***qldwater***. The workshop attendees agreed upon a list of practices that they felt would be useful for all SPs regardless of size, region or other factors. It was also decided, however, that some smaller SPs may face constraints that lessened the possibility of implementing some practices.

The recommendation of the workshop was that the identified list of QC Practices be provided to all Queensland SPs in a format that allowed the immediate adoption of these practices wherever possible. It is based on this recommendation that this document has been developed. The workshop further recommended that a template be provided to facilitate use of QC Practices. A draft template and details on its use are outlined in the next section.

Overview: How to use this document

This document contains a series of Quality Control Practices that can be implemented by Queensland SPs in order to improve the reliability and therefore usability of data. Each practice is listed in the contents page which directs the reader to the relevant section. Each practice has its own section, which briefly outlines the rationale and use of the practice, with examples and instructions on use of the accompanying template where relevant. This document has been developed with the aim of enabling immediate adoption of the data quality control practices that are considered appropriate by Service Providers. As most of the checks require no further technology, just some time and knowledge, it is considered that most of them are feasible to be adopted by all SPs.

The QC Template

This document is accompanied by a generic template that can be used for any data QC checks (including any not listed in this document). A set of specific templates has also been provided to facilitate some of the QC checks detailed below. The use of these specific templates is explained in each of the relevant sections (pp. 5 - 11).

A Note for SWIM Users

SWIM is the Statewide Water Information Management System. Much of the data that SPs store and report is submitted via SWIM. From this year onwards, SWIM will have built-in Quality Control Checks. More specifically, SWIM will generate a report that will be provided to users highlighting each of their data that has failed one or more SWIM quality control checks (which are a subset of those described in this document). This SWIM QA report will make data quality control simple and efficient for SWIM data. However the template provided here, in combination with the information in this document, can be used for any data collected by SPs or could be used to perform quality control in addition to that performed automatically for data submitted to SWIM.

How to Check Data

If any data fails a particular QC Check then it should be double-checked for any errors. This process may involve checking a spreadsheet into which the value was initially entered to ensure the final value is the same as the original, or contacting another staff member who is responsible for recording that particular data to check the original value. If the data is found to be incorrect, then it should be changed immediately (in all locations that it is stored).

If after investigation any data that failed are found to be correct, then this may support the need to investigate why the data are unusual. If the values for a certain data point are consistently extreme compared with expected ranges, this may represent a need for change in the process that the data is measuring. It may also mean that the measuring equipment or procedure for creating the data is flawed, and may need to be replaced or redesigned, or staff may need to be retrained. It may be the case that a value is particularly high or low because an old meter needs to be replaced, or that staff members have been entering data using incorrect units. Finally, some extreme values may be legitimate, and despite failing a QC check, may represent the actual situation correctly. If this occurs frequently, the

QC rule itself may need to be modified – caution should be taken to be very certain about the correctness of data and the need to change a QC rule before any rules are modified.

What it could mean if your data is failing QC Checks

- Extreme values represent a unique situation for a particular location
- equipment for recording data is broken or in need or recalibration
- Data has been recorded using the wrong units
- Human error in copying data from one data storage to another
- Human error in recording data
- An automated process for recording, transferring or aggregating data has design flaws or bugs
- The definition or format requirement of a data point has been misinterpreted
- A variety of other problems

QC Practices: Descriptions, Instructions and Examples

1. Comparison to historical data

Description

To carry out this check, current values are compared with values for the same data point and scheme for a previous time period using a statistical rule for acceptable fluctuation. This comparison makes the assumption that data will not change by too much from year to year. An acceptable range of change should be decided for each data point depending on the type of process being measured (e.g. +/-25%), and any data point that changes by more than this acceptable range should be double-checked and investigated. It is of course possible that a datum has, for one reason or another, changed to the extent that it exceeds the arbitrary acceptable range. Even so, all data that fall outside the acceptable range should be checked, and a legitimate reason for the change identified where possible.

For all SWIM Indicators, appropriate ranges have been determined in consultation with Queensland SPs, and the results of this QC check are automatically included in the report provided to SWIM users. As an example, the SWIM indicator CS4 (Total Connected Properties – Water Supply) is submitted to the following check: this year's value = last year's value +/- 25%. This rule means that the total connected properties in a SP or scheme are not expected to increase or decrease by more than 25%. Lower or higher variances could be used locally depending on known growth rates of different communities.

Use of QC Template

QC Check:	Comparison to Historical Data
Date Performed:	24/05/2011

Datum	Site	Current Value	Units	Acceptable Change	Previous Value	Actual Change	Outcome
e.g. SWIM Indicator title, or any other data point	*The site of the recorded data e.g. WSP-wide*	*The value recorded at*	*E.g. ML, no. of connections, $000s*	*Range of change acceptable e.g. +/- 25%, +/-1*	*Value for previous time period e.g. Last year*	*The calculated actual change*	*Pass or fail*
CS4	WSP-wide	19.2	1000s of properties	+/- 25%	17.4	10.34%	Pass
CS4	Scheme 1	7.8	1000s of properties	+/- 25%	5.8	34.48%	Fail
CS4	Scheme 2	1.3	1000s of properties	+/- 25%	1.3	0.00%	Pass
CS4	Scheme 3	10.1	1000s of properties	+/- 25%	10.3	-1.94%	Pass

Figure 2: Example of QC Template

The example in Figure 2 uses CS4, the SWIM Indicator discussed above. In the above example, the change from one year to the next in Scheme 2 and Scheme 3, as well as SP-wide, is between -25% and +25%, therefore these 3 data have passed the test. On the other hand, the change in scheme 1 from 5,800 connected properties to 7,800 connected properties is an increase of over 34%. Therefore, this datum fails the QC Check. If a legitimate reason for such a large increase is known (for example a large new development), then this increase may be correct, otherwise it is probably an error. Either way, the value should be double-checked.

2. Comparison to other schemes and SPs

Description

This QC Check requires some information beyond your own SP's data. After data is reported, various state and commonwealth agencies collate and analyse the data and make it available in various formats. For SPs in Queensland, the relevant annual reports are the Comparative Report produced by ***qldwater*** using data from SWIM, and the NPR that uses data from largeSPs across Australia. The information found in these reports can be used to compare your own SP and schemes to others around Queensland and Australia. If any data are found to be very different to state and national values, then these data may be incorrect and should be double-checked or investigated.

A good way to establish if your SP's data are extreme or unusual is to compare your SP's data to percentiles and the median where possible. While it is possible that extreme data are accurate, if they are too far from the median, or are consistently below the 10h percentile or above the 90^{th}, the data should be checked for mistakes. A common mistake is to enter values in the incorrect unit. For example, if the SWIM indicator CS1 (Population receiving water supply services) is entered in whole numbers instead thousands, it will likely be above the 90^{th} percentile (see figure 4 below).

The SWIM Comparative Report provides data in tabular and graphical formats for efficient comparisons with other schemes. The Report will show the values for each indicator for each scheme compared to the medians and 10^{th}& 90^{th} percentiles for other Queensland schemes of the same size, same region and same soil type. The graphs and tables are presented in a meaningful way, allowing meaningful interpretation (of data that has already been reported) without requiring further analysis. For quality control of data yet to be reported, the main use of the comparative report would be to provide a summary of probable ranges for data.

Use of QC Template

It would certainly be possible to simply copy a set of data (from NPR or SWIM Comparative Report) and eyeball this data against your own. However, the example used here will compare against a statistical summary of other schemes and SPs. This is the method used in the SWIM Quality Control Report, which provides a comparison to the 10^{th} percentile, median, and 90^{th} percentile of the current year's data and the previous year's data from across the state.

QC Check:	Comparison to Statistical Summary				
Date Performed:	10/06/2011				
Datum	Current Value	Units	Statistical Summary of Comparison Data		Outcome
e.g. SWIM Indicator title, or any other data point	*The value recorded at present*	*E.g. ML, no. of connections, $000s*	*Median*	*90th Percentile*	*Pass or fail*
WA16	19133580	ML/year	1007	22453.44	Fail
WA17	1050	ML/year	0	3161	Pass

Figure 3: Example of QC Template (some columns hidden in this example for readability)

The example in Figure 3 shows Volume Residential and non-trade (only) waste sewage collected (WA16) and Volume Trade waste (only) sewage collected (WA17) with statistical summary data from ***qldwater***'s Comparative Report. In this example, the value for WA17 falls between the median and the 90^{th} percentile for all Queensland Service Providers. Conversely, the value for WA16 is much higher than the 90^{th} percentile, and therefore fails this QC check. As the value is so high, it is likely that it has been recorded in kL/year instead of ML/year. Assuming this to be the case, the corrected value of 19133.58 would fall below the 90^{th} percentile and pass the QC check. However this assumption should not be relied upon, and the value should still be checked rigorously.

3. Relational Data Checks

Description

Relational Data Checking is checking that data makes sense as a set. As many of the data are measuring things in the real world that are related, it makes sense that relationships also exist within the data. For example, it would not make sense for a SP to have a total annual expenditure of $10,000,000 and have spent $15,000,000 on infrastructure upgrades. While many of the relationships may be less obvious than this example, some very clear relationships do exist, and can be defined using simple formulae. These formulae can then be used to check that your recorded data satisfies the logical relationships.

One means of relational data checking is checking water balances following the basic rule that *water sourced = water used/lost.* This is applicable to schemes and SP-wide values. For example, total water sourced = total water consumed + losses (allowing for a small error margin eg. +/- 5%). Using SWIM Indicators, this rule can be stated as WA48 = (WA11 + WA46 + WA13) +/- 5%. By simply adding the recorded values for WA11, WA46 and WA13 and comparing to the recorded value of WA48, errors can be very easily spotted.

There are also other Relational Data Checks besides water balances. Examples include sewerage effluent treatment level: percentage treated to primary + secondary + tertiary = 100%); water pricing typical bill = average consumption times appropriate block tariff fees plus fixed fee.

4. Derived Values

Description

While some data are measured individually and recorded by users, other data are just derived from values already recorded. For example, in SWIM the volume of residential and non-trade waste sewage collected (WA16) and the volume of trade waste sewage collected (WA17) are entered by SPs. The total volume of sewage collected (WA18) is then calculated by adding the two entered values together. This process can be represented mathematically as WA18 = WA16+ WA17.

While using SWIM templates, derived values update automatically when one of the values they are dependent on is updated. For example, if WA16 is increased by 200, WA18 will automatically increase by 200. However, if not using the SWIM templates, or if the derived values have been manually written in the SWIM template, then these automatic updates will not occur. As such, it is worthwhile to use a QC Check to ensure that derived values are correct, relative to the values they are dependent on. To do this, a list of derived values can be made, along with the formulae used to derive them, then these calculations can be carried out manually (i.e. in Excel) to check the recorded value is correct.

Use of QC Template

QC Check:	Derived Values Checking					
Date Performed:	25/05/2011					
Datum	Site	Current Value	Units	Derivation Formula	Manual Calculation	Outcome
e.g. SWIM Indicator title, or any other data point	*The site of the recorded data e.g. WSP-wide*	*The value recorded at present*	*E.g. ML, no. of connections, $000s*	*E.g. Sum of Schemes Values, WA18 = WA16 + WA17*	*Manually Calculated Value using derivation formula*	*Pass or fail*
WA16	WSP-wide	19133.58	ML	N/A	N/A	N/A
WA17	WSP-wide	1050	ML	N/A	N/A	N/A
WA18	WSP-wide	20183.58	ML	WA18 = WA16 + WA17	20183.58	Pass
CS8	WSP-wide	61.77	1000s of connections	N/A	N/A	N/A
WA19	WSP-wide	1420.62	kL/connection	WA19 = WA18/CS8	326.753764	Fail

Figure 4: Example of QC Template

The example used in the description above is illustrated in Figure 4. The recorded values for WA16, WA17 and WA18 have been copied from the spreadsheet in which they are stored. The value for WA18 is then calculated manually from the recorded values for WA16 and WA17. In this example, the manually recorded value is the same as the recorded value, so this datum passes the QC check.

On the other hand, WA19 (Average volume sewage collected per property) has failed the QC check, as the recorded value is significantly different to the calculated value. This may simply mean that the recorded value needs to be updated to reflect a change in CS8 (Total connected properties – sewerage) or another value on which WA19 is dependent. It may also indicate that an error has been made in one of these indicators, and all of them should be double-checked.

5. Numeric/text format

Description

Data is usually required to be stored in a certain way in order to be meaningful. A SP should be aware of the format in which each value is required to be stored. Using this information, a list of indictors that are required to be purely numeric can be generated, and the recorded values of all these data can be checked to ensure they contain no text, and only numbers. Common mistakes include entering $ for dollar values or % for percentages, or simply writing notes about a value in a numeric field rather than in a dedicated 'comments' field. This can skew analysis of the data by the service provider at a future date. For SWIM data, the QA Report will alert users to any numeric fields that contain characters that are not allowed. To avoid confusion, SWIM has been created so that there are no instances in which text should appear in a numeric field. If a value is found that fails this QC check, it must always be changed.

Use of QC Template

QC Check:	Numeric Format			
Date Performed:	25/05/2011			
Datum	**Site**	**Current Value**	**Units**	**Outcome**
e.g. SWIM Indicator title, or any other data point	*The site of the recorded data e.g. WSP-wide*	*The value recorded at present*	*E.g. ML, no. of connections, $000s*	*Pass or fail*
EN9	WSP-wide	48.61	tonnes CO2/1000 connections	Pass
EN10	WSP-wide	312.84	tonnes CO2/1000 connections	Pass
EN11	WSP-wide	95.7t	tonnes CO2/1000 connections	Fail
EN12	WSP-wide	#VALUE!	tonnes CO2/1000 connections	Fail

Figure 5: Example of QC Template

The example in Figure 5 uses the SWIM indicators EN9, EN10 and EN11, which measure greenhouse gas emissions from water, sewage and other sources respectively. The units for these indicators are tonnes of CO2 per 1000 properties, and the value is to be recorded as purely numeric. In the above example, EN11 has been recorded incorrectly, with a 't' in the value field to denote the units. Therefore this data point fails the QC check and should be changed.

Note that EN12 (total greenhouse emissions per 1000 properties) is a derived indicator calculated by the addition of EN9, EN10 and EN11. Due to the error in EN11, EN12 also fails as it is unable to be calculated. This highlights the importance of ensuring that numeric data contain only numbers; so that further calculation and analysis can be performed on the data.

6. Acceptable Ranges

Description

Some data are defined in such a way as they are limited to a certain range of values. If this is the case, recorded values should be checked against these acceptable ranges to ensure that they do not fall outside them. For example, the SWIM indicator WA27 measures the percentage of effluent that is recycled. The unit for this indicator is percent (%), therefore the value must be between 0 (no effluent recycled) and 100 (all effluent recycled). Any value that falls outside this range cannot be correct by definition, and must therefore be the result of an error and be changed. This example is illustrated in Figure 6.

Use of QC Template

QC Check:	Acceptable Ranges
Date Performed:	25/05/2011

Datum	Site	Current Value	Units	Acceptable Range	Outcome
e.g. SWIM Indicator title, or any other data point	*The site of the recorded data e.g. WSP-wide*	*The value recorded at present*	*E.g. ML, no. of connections, $000s*	*E.g. 0 - 100, 0 - 1*	*Pass or fail*
WA27	WSP-wide	23.14	%	0 - 100	Pass

Figure 6: Example of QC Template

7. Ineligible Zero Values

Description

For some data, it is impossible or very unlikely the correct value is 0. A list of these data can be compiled by SPs, then the recorded values checked to ensure that they do not violate this rule. For example, SWIM indicator CS1 represents the Population receiving water supply services. It is impossible that a Water Service Provider could have a value of 0 for this indicator. Conversely, the total volume of water taken from desalination (WA3) may be 0 for some SPs.

Use of QC Template

QC Check:	Ineligible zeros				
Date Performed:	25/05/2011				
Datum	Site	Current Value	Units	Outcome	
e.g. SWIM Indicator title, or any other data point	*The site of the recorded data e.g. WSP-wide*	*The value recorded at present*	*E.g. ML, no. of connections, $000s*	*Pass or fail*	
CS1	Scheme 1	1.3	1,000s	Pass	
CS2	Scheme 1	0.8	1,000s	Pass	
CS3	Scheme 1	1	1,000s	Pass	
CS4	Scheme 1	1.8	1,000s	Pass	
CS1	Scheme 2	0	1,000s	Fail	
CS2	Scheme 2	1.2	1,000s	Pass	
CS3	Scheme 2	1.1	1,000s	Pass	
CS4	Scheme 2	2.3	1,000s	Pass	

Figure 7: Example of QC Template

In the template for ineligible zeros, SPs should only list data for which 0 is ineligible (which may differ between SPs). In Figure 7, the user has listed CS1-CS4 for 2 schemes, and it is clear that CS1 has failed the QC check. This QC check usually fails when there is no data for a certain point, which should be recorded as ND (No data). If this is mistakenly recorded as 0 or left blank and subsequently interpreted as 0 by a piece of software, this datum will fail the QC check. It is important to remember that 0 has a specific meaning and represents a known measured value of nothing, which is not the same as having no data, or a particular measurement not being relevant.

Missing Values

The inverse of this error is when a value that should be recorded as 0 is mistakenly left blank or recorded as ND. Once again, it is important to remember that if a value is known to be 0, this is very different from having no data.

8. Validation

Description

Validation is the use of a second independent measuring technique to check the reliability of a first measurement. This type of quality control is very valuable, as a second measurement confirming the first, highly increases the confidence that this measurement is correct. At the same time, it is not always possible to measure all types of data using different methods, and it may not always be feasible for SPs with less resources to use this quality control technique.

9. QC checks in sampler/device

Description

For some SPs, it is feasible to use sophisticated data sampling technology with data-checking capability. A fairly common example in use in Queensland is a handheld PDA device used for reading meters to measure water usage at properties. These devices may include QC checks to compare to previous values and alert users if there is an unexpected value. While these sophisticated devices represent an excellent quality control, they may not be affordable or practical for all SPs. At the same time, it is important to remember that using technological methods of sampling or recording data does not eliminate the need for further QC checks before using or reporting data. There is still the chance that errors are made when data is transferred or aggregated, and QC checks should always be used to ensure the reliability of data.

10. "Eyeballing"

Description

Eyeballing is informally judging data against previous knowledge of an acceptable value or a 'gut feeling' about a certain data point. This type of quality control is cheap and efficient but not particularly rigorous. It is widely practised in a very informal way. Commonly, the officer responsible for gathering and preparing data for reporting will examine the data they are entering and make sure there is nothing that stands out as obviously incorrect. This method is extremely valuable as a first check of data and to ensure there are no obvious problems, but is insufficient to provide certainty of data quality assurance.

Using the QC Template

QC Check:				
Date Performed:	24/05/2011			
Datum	Site	Current Value	Units	Outcome
e.g. SWIM Indicator title, or any other data point	*The site of the recorded data e.g. WSP-wide*	*The value recorded at present*	*E.g. ML, no. of connections, $000s*	*Pass or fail*
WA1	WSP-wide	160	ML	Pass
WA2	WSP-wide	3	ML	Pass
WA3	WSP-wide	0	ML	Fail

Figure 8: Example of QC Template

As eyeballing is an informal quality control check, the generic template can be used (if a template is needed at all). In the example in Figure 8, a set of data has simply been copied into the generic template. The user must have knowledge of the indicators to be able to determine that WA1 and WA2 have passed, whereas WA3 has failed. This simply means that the user believes that WA3 may be incorrect and intends to double-check it.

SP Practices to Increase the Value of QC

The QC checks described above are useful tools in checking the correctness of data. However, to be effective, they must be incorporated to the day-to-day operational procedures within the SP, and be supported, for example, with training for data coordinators/managers, standard instructions manuals and dictionaries, and access to further information for solving problems when they arise.

Institutionalised Quality Control Protocols

Documenting a set of standard procedures for use by the entire SP will greatly improve the outcomes of any required practice, including data quality control. Documentation ensures that correct procedures are able to be found at any time, regardless of whether or not a staff member who already has that knowledge happens to be present. Furthermore, a documented standard procedure ensures that there is an official record of the correct way to do things. This means that throughout the organisation every staff member will be using the same procedures. This is important for all steps in the data cycle, from collection to quality control.

The other important benefit of documenting a set of protocols is that the knowledge won't be lost when staff members move on to other organisations or retire. A documented protocol also assists in training new staff. Ensuring consistency of data handling across time is vital for reliability of data.

In order to 'institutionalise' the quality control protocols, a set of data management procedures & records should be easily accessible by all data coordinators and users. These can take any format, but generally should include some standard information (listed below). In most cases, it will be most efficient to develop these documents for groups of indicators rather than for each indicator; for example for the water consumption indicators; the water sources indicators.

- Relevant indicators
- Sources of data (locations in the organisation is the data generated)
- Method used to generate data for each indicator (e.g. meters, calculations (give details), laboratory measurements (give details)), include the frequency of data collection (note if different methods are used for different locations, identify these)
- Uses (e.g. Which internal/external reports)
- QC practices to use (e.g. select from the QC practices listed above)
- Record of QC assessments:
    - Which (batches of) data (e.g. identify by dates and location)
    - Date of QC assessment
    - Which QC rules were tested and results (pass/fail)
    - Action taken in response to fails

Data Management & QC Training

Training staff in data management and quality control greatly improves the reliability of data. Passing on knowledge and experience to new staff members can significantly reduce the amount of time and effort required by these staff members to learn the procedures in the organisation. It also has similar benefits to documentation; it creates consistency across the organisation as well as consistency over time. Training staff in data management and quality control also provides an opportunity to reinforce for staff that these practices are very important for the organisation.

Data Dictionary/Thesaurus

A ‘data dictionary’ is a record of definitions of terms that are relevant to data. At the workshop that was held prior to the development of this document, attendees agrees that a data dictionary would be useful to ensure that all SPs and other organisations with an interest in water services were using words to mean the same thing. Once again, this promotes consistency not just throughout the organisation, but throughout the entire state.

Users forum to exchange data knowledge

Another initiative that was strongly supported by the workshop attendees was the idea of a users forum to identify common challenges, problems, and most importantly, solutions. Such a forum could potentially increase the reliability of data and quality control by allowing users to share valuable information with each other, so that innovations and developments would not be confined to the SP in which they were made.

Different Data Categories and the Relevant QC Checks

To assist in using the Quality Control practices described above, we have identified suitable QC checks for groups of SWIM Indicators. The figures below can be used to determine which QC checks are suitable for which SWIM data. In addition, users can extrapolate from these tables to establish suitable QC checks for other data besides that submitted to SWIM.

Table 1 lists different categories of information that data measure. Along with each category are listed some example SWIM indicators, some relevant QC checks and notes on their use.

Table 2 lists some specific data types, and notes on QC checks that are relevant to that type of data.

Please note that these two tables are to be used together, and that all data should be considered for all QC checks; do not use the QC checks just from one table or the other.

Table 1: Grouped by data category

Data Category		Example Indicators	Comparison to historical data	Relational Data Checks	Ineligible zero values
Water and Sewerage Volumes	WA1	The total volume of water taken from surface water	Water volumes are not expected to change much from year to year, therefore the **+/-25%** rule is used	There should be relationships between volume indicators. Along with the generic relationships used by SWIM templates, your SP may be able to use specialised knowledge to define expected relationships.	Water volumes would not be expected to contain 0 values. Most SPs should have some volume for all water indicators.
	WA2	The total volume of water taken from groundwater			
	WA3	The total volume of water taken from desalination			
	WA18	Total volume sewage collected	Total and average sewage volumes are similarly expected to fluctuate little, so the **+/25%** rule is used for these also		Similarly, sewage volumes would not be expected to contain 0 values.
	WA19	Average volume sewage collected per property			
Water and Sewerage Volumes for occasional or new users	WA27	Recycled water (percent of effluent recycled)	If your SP is establishing new recycled water operations, it would make sense that these volumes could change by more than +/-25%. For this indicator +/-100% is used.	Relationships between volume indicators should still exist for occasional or new users. This QC check only compares data to other data from teh same reporting period, so occasional use or non-use will not make a difference.	Zero values are possible for occasionally used water sources or other similar data
	WA26	Total volume recycled water supplied (within the WSPs geographic area of responsibility)	These 2 indicators may reasonably show even more fluctuation so +/-200% is used.		
	WA20	Volume of recycled water supplied: Residential			
	WA1	The total volume of water taken from surface water	The volumes for these indicators are not expected to change much so +/-25% is used by SWIM. However, for certain schemes, these values may fluctuate more than this when a water source (eg bore) is only used occasionally.		
	WA2	The total volume of water taken from groundwater			
	WA3	The total volume of water taken from desalination			
Compliance data	HL3	% of total population where microbiological compliance was achieved	Compliance should not change much from year to year as they are expected to be based on long-term policies. SWIM uses +/-20% to compare these indicators to historical data.	It may be possible to define relationships for these data.	Compliance data should never contain 0 values or missing data.
	EN4	Percent of sewage volume treated that was compliant			
Descriptive Data	FN31	Asset valuation methodology	These data are text indicators. As such, none of these QC checks can be applied. The data can still be "eyeballed" or double-checked to increase reliability.		
	PR1	Water pricing: Tariff Structure (description)			
	PR5	Fixed charge - water: description of basis of charge			
Pricing Data	PR3	Fixed Charge - water: value	These data are unlikely to change from year to year anf the +/-25% rule is used.	Relational Data Checks can be defined for pricing data. For example there is a relationship between the fixed charge and	Zero values are not expected for pricing data.
	PR6	Usage Charge 1st Step: value of charge			
	PR43	Annual bill based on 200kL/a - water			

Table 2: Grouped by QC Checks

QC Check	Notes on Application
Comparison to other schemes and SPs	All numeric indicators can be compared to values from other WSPs and other schemes. For QC purposes, a good method is to compare your WSP or scheme's value to the median and 10th/90th percentiles of other WSPs. This can highlight extreme values that may be incorrect or recorded using incorrect units.
Derived Values	Within every data category, some indicators are measured and others are derived. No matter what the category, derived values can always be checked by performing a separate manual calculation.
Numeric/text format	Within every data category, some indicators are required to be entered as purely numeric, others allow text. No matter what the category, values can always be double-checked to ensure that numeric data contains only numbers.
Acceptable Ranges	Knowledge and experience, along with analysis of previous data, may allow useful acceptable ranges to be defined. This process can be conducted for all numeric data to produce a specific acceptable range for each one.
Validation	The usefulness and feasibility of validation is largley dependent upon your SP's capabilities and specific daum being measured. Validation requires a secondary independent measurement, and a generic rule for when this technique can be applied is impossible to determine.
QC checks in sampler/device	Similar to validation, the possibility of using this method is determined by the SP's specific situation. The most common use in Queensland is PDAs used to record meter readings that contain built-in QC checks.
Eyeballing	This is an informal checking method rather than a rigorous QC check. It is possible with all data at all stages of measurement, recording, processing and reporting.

Acknowledgments

qldwater would like to thank those representatives from SPs who participated in the survey and attended the workshop that led to the creation of this document. They would also like to thank BoM who provided the funding for the completion of this project. Finally they would like to thank all members of ***qldwater*** for their ongoing support.