DQM Information Paper 8.2.3
"Good Data" Made Tangible

By Revital Katznelson, Ph.D

1.0 About this Information Paper

(This paragraph is essentially common to DQM Information Papers from the 8 and 9 series. If you have seen it already, please skip to the next paragraph below). This Information Paper is a new type of guidance. It has been created for our new integrated system of guidance and tools for water quality monitoring called "the Data Quality Management (DQM) System". DQM is implemented by the Clean Water Team (CWT) where needed to support collection of **reliable** data of **known quality** in a fully documented, **scientifically defensible** manner. Most DQM materials are delivered in Parameter-Specific Folders, which provide both the traditional "protocol" materials and new, expanded guidance in three types of inter-related documents: Fact Sheet, Information Paper, and Standard Operation Procedures.

This Information Paper (IP), a part of the generic DQM contents materials, provides "big picture" information on data quality. The IP also points the reader to further DQM guidance that provide more details or give instructions on the "how to" mechanistics of assuring, maintaining, and communicating data quality. If you are a Trainer or a Technical Leader of any monitoring project, this may help you separate the different aspects of data quality and find out what it takes to communicate how good your data are, and what they represent in the environment.

Section 2 of this IP lists the different attributes of good data. Section 3 provides an overview and a brief description of the process of gathering good data. Communication of data quality is discussed in Section 4. Finally, the "Sources & Resources" section (Section 5) provides contact and website leads into further information.

2.0 Attributes of Good Data

Box 8.2.3-1 shows a breakdown of different aspects and attributes commonly mentioned about environmental monitoring data. The key word here is: **Different**; each provides a different type of information about the data and has to be communicated separately. Each of these aspects and attributes has more than one dimension, and together they form a multidimensional matrix that can be "sliced" in multiple ways. The guidance literature is teeming with frameworks and methodologies that translate all these aspects into operations, or a process, or a philosophy, or even a dogma. This Information Paper was designed to relate to the "project planning action items" aspects as it walks you through the data collection process, and to the communication language you might use to communicate the outcomes of the process.

Box 8.2.3-1: Attributes of GOOD DATA

USABILITY of the data –

* Capable of answering questions and supporting management decisions
* Complete Parameter Package: all supporting Characteristics (Analytes, WQ parameters, etc.) included
* Comparable to other (defined) data sets in terms of sampling design & procedures
* Adequate Measurement Quality Objectives in terms of sensitivity and tolerated error
* Adequate Statistical Power of the dataset (e.g., number of samples) and adequate Confidence Level
* Scientifically defensible, including in court
* Collected by operators for whom there is no prejudice
* Reported in formats that can be easily read and transformed by others

RELIABILITY (Credibility) of the data –

1. Known probability that the reported value indeed falls within the range of error specified for it, and
2. Complete supporting documentation is provided (e.g., location, methodology, quality checks), and
3. Honest reporting by field operators

VALIDITY of the data –
* The test or analysis used to collect the data was shown to be valid; the instrument was functional (i.e., the measurement system met its performance criteria)

…ACCEPTABILITY (or COMPLIANCE) of the data –
* Agreement with data quality objectives, quality check requirements, and acceptance criteria has been confirmed

KNOWN QUALITY of the result: Three DISTINCT aspects -:

First: Quality of the measurement itself (accuracy, precision, detection limit, resolution)
Second: Sample integrity (lack of deterioration, lack of contamination)
Third: Representativeness of the measured value (how does it represent “true” environmental conditions, across time & space)

CORRECTNESS of the data –

- Consistent interpretation of "menu options" for verbal or numeric range categories
- Correct transfer of information from observer to "scribe"
- Correct recording, copying, and data entry into electronic formats

CERTIFICATION –

- Laboratory analyses were made by a certified laboratory
- Laboratory work was checked against certified standards
- Field work was done by certified operators, survey work done by certified surveyors
- Field work done with or calibrated against certified instruments

3.0 Process Overview

This section walks through the four major items of the data generation process, namely, planning, developing quality objectives, monitoring and checking the quality of the measurements, and validating the data. Gathering of good environmental monitoring data begins with good **planning**. If you have a clear monitoring objective use it to formulate a very specific study question, it will be easier to define the proper **objectives** and design the study accordingly so your data will answer the question and thus will be usable in that respect. Well-thought plans help with assuring and tracking the performance of the data **collection** effort itself, and implementation of appropriate data quality management tools can greatly facilitate the data verification and data **validation** steps. The following sections elaborate on these four milestones.

3.1 Planning: question, parameter package, comparability, and sampling design & procedures

Because very few societies or organizations can afford and justify collection of data for the sake of collecting data, the most important aspect of data usability is the capability of answering questions and/or supporting management decisions. This leans on a well-defined question, or decision to be made, and on adequate breadth of complementary information. It means that you need to choose good indicators and to plan for a good package of characteristics (a.k.a water quality parameters) to measure. For example, if your question is something like "what is the incidence of ammonia toxicity in the Sycamore Creek watershed" and you have multiple samples analyzed for ammonia, you will not be able to use the data if you do not know the ambient pH at the time each sample was collected because you need the pH to calculate the un-ionized (and toxic) fraction of ammonia. Similarly, if you collect dissolved copper data for comparison with water quality benchmarks (i.e., with a concentration thought to be not toxic to aquatic life) you will not be able to use the data if you do not know the hardness of the same sample waters, because the hardness affects the toxicity of copper.

A classical field measurement example is about someone trying to interpret the ecological significance of 7.0 mg/l dissolved oxygen (DO) without any information about flow: this value may indicate very healthy conditions if measured in a non-flowing (stagnant) waterway, but may raise concern if recorded in a fast-flowing stream. Needless to say, you cannot interpret any DO results without knowing the temperature!

Statements about comparability of data must always have a "to what" clause included, but need to also include which attributes can be compared. For example, your data will be comparable to the Surface Water Ambient Monitoring Program (SWAMP) data if you choose your sampling location in the "directed" sampling design principle and collect the sample in the centroid of the flow at the depth of 0.1 m below the surface. In California streams, data collected during dry weather are very different from data collected (at the same spot) during storm runoff flow. You can compare dry weather data in your stream to dry weather data in another stream, and even to storm runoff data in your stream, but not to storm runoff data in another stream. Thus, comparability is tightly related to representativeness, i.e., what each Result represents in the environment. Defining the types of Stations you will monitor (a stream or an outfall?), the sampling design principle you will use (random or knowledge-based?), the conditions (wet or dry? Worst-case or ambient?), the sampling procedures you will use, etc., is an essential part of the planning step that will determine how comparable your data are (and to what!). Determining how good your data need to be in terms of error and statistical power is another aspect of comparability, described below.

3.2 Quality Objectives: tolerated error and statistical power

Most people think about accuracy and precision when they hear the term ”good scientific data”. However the fact that you do not need highly accurate and precise data to answer most monitoring questions is often overlooked. The key word here is: **Adequate.** And the planning consideration is: how much error can we tolerate and still be able to answer the question. For example, if you plan to use your data for the “source identification” chapters of your nutrient TMDL, you can tolerate large error (e.g., 100 or 200%, because it does not matter whether this tributary had 3.4 mg/l or 3.8 mg/l of Nitrate; what matters is that it was significantly higher than the other tributary which had 0.4 mg/l of Nitrate, so now you know which tributary is a major source). But if you are conducting the “loads assessment” part of your TMDL you need much tighter data (e.g., no more than 10-20% error) because you will use the data to calculate the loads coming from each tributary.

Another aspect of data usability has to do with having a dataset with adequate statistical power, i.e., enough samples, to enable meaningful statistical comparisons that will yield significant results. For example, if you want to compare the E. coli counts before and after implementing your Management Measures (e.g., fencing the stream channel to keep the cattle out), you will need to count enough samples at each phase to show a significant reduction. The number of samples will greatly depend on the ‘natural’ variability in E. coli concentrations.

USEPA has developed a process called “the data quality objectives process”, or DQO process, which walks planners through six step that assist with problem definition, question/decision formulation, definition of spatial and temporal scale, and determination of tolerable error and desired confidence level (USEPA 2000, QA/G-4). It is highly suitable to situations where the user has to characterize a given parcel of land (or a waterbody) through an average that represents the entire unit. It assumes that you already know the distribution of an analyte, or the inherent variability of the characteristic of interest, within your ‘parcel’. The DQO process can be also used in the design of effectiveness monitoring studies and in many other situations. However, the application of this framework has to be done judiciously so please seek further guidance.

3.3 Monitoring activities

Many aspects of data usability and comparability relate to the “on-the-ground-operations” during the data collection phase. Known quality is an important one: most data users will be very reluctant to use or base decisions on data of unknown quality. In a nutshell, what needs to be known is the performance of the Measurement System (defined as the entire process of sampling

& analysis or field measurements) in terms of detection limit & resolution, accuracy/precision error, and sample integrity. For this you can implement is a set of tools that help you generate data of known quality. The following steps provide an example of what it takes, for field measurements:

1. If you calibrated an instrument, collected data, and now you are ready to calibrate again, do an "accuracy check" first and **record the reading before any calibration adjustments**. [this is the same as "post-calibration" check].
2. Run periodic **accuracy checks** to all your non-adjustable instruments
3. **Repeat discrete field measurements** with each Instrument at least twice on every Trip
4. **Write** it all down, preferably with Instrument ID.

It all boils down to three actions, namely
-- Assign a unique **Instrument ID** to every measurement device
-- Link every Result with the Instrument that was used to measure it
-- Link every batch of Results with Instrument calibration and **accuracy checks** records, and Instrument **repeated measurement** records, for a given period of time

Any procedure that assures and controls the quality of your measurements will help, particularly if it has clear direction on how to affect, check, record, and report data quality. For example, if you are measuring specific conductivity with a meter you will need to **affect** its **accuracy** (e.g., adjust instrument calibration often), **check** its accuracy (e.g., the reading in the Standard before next adjustment), **record** (e.g., the value of your Standard and what your instrument reads in it), and **report** its accuracy (e.g., how much it had actually drifted from the calibrated state). Similarly, you will **affect** your **precision** by being consistent, **check** it by conducting repeated measurements, **record** results of these repeated measurements, and **report** your precision as the relative percent difference (RPD). The DQM guidance shows you what to capture and how to calculate the accuracy and precision of a variety of field instruments (e.g., see DQM-SOP-3.1.3.1 for instructions on the pocket conductivity meter).

So far we have talked about the measurement error, i.e., the "noise" associated with the measurement system. What your result represents in the environment is totally separate from this Noise, and depends on "where" and "when" you conduct your field measurements or collect your samples. It is the duty of all Project personnel to apply the plans and the logistics and to assure that the monitoring Results represent what you intend them to represent. If you read DQM-IP-8.2.4 on representativeness you will realize how much control you have on the process and how many choices you actually have.

There are many other aspects of good data that manifest themselves during the process of data collection, most of them are self-evident in Box 8.2.3-1, and some are elaborated upon in other DQM guidance (see section 5 below). But what about reliability? There is only one aspect of data quality, of the entire list in box 8.2.3-1, that relies on an assumption of faith: that operators report honestly. This is the only thing that cannot really be proved. For everything else, there can (and should) be documentation!

3.4 Data Verification and Validation

The quality of data cannot be known if the basic documentation about measurement quality and representativeness has not been captured, but even if they were, they may need to be "processed", compared to performance criteria or requirements, and communicated. In other words, the validity of the data has to be established, with adequate documentation. The data validation and verification process can be simple or complex, depending on the magnitude or scale of the monitoring effort, but either way it involves a set of defined steps (e.g., USEPA 2002, QA/G-8). If you are using DQM tools, use the instructions in DQM-SOP-9.3.2.2 (in the CWT Toolbox) for error calculation and data validation.

4.0 Communicating Data Quality

4.1 Words to use

The hardest part of communicating data quality is finding a dictionary that clearly defines the appropriate terms. If you had an opportunity to peruse the glossaries of the USEPA Quality System or the USGS guidance, you will find a rich selection of words that mean the same thing, and you might find words that mean more than one thing too. The National Water Quality Monitoring Council (NWQMC) has compiled a long list of words that have multiple definitions, but the monitoring community is still awaiting a decision which definition will be used by everyone. In California, different agencies and groups use their own glossaries. The Clean Water Team glossary contains definition of data quality terms as used in the DQM guidance, and it is periodically aligned with the SWAMP glossary to assure comparability.

4.2 What to communicate

An important aspect of communicating data quality, and making data usable, is provision of descriptors that go with the data, also known as metadata, or 'data about the data'. The DQM contains a comprehensive list of "information bits" that can be captured and communicated to describe the monitoring results. Some of these bits are communicated within the data collection effort (for example, the calibration records that are used to calculate accuracy), other bits might go beyond the Project, into a report or a website or a central database, along with the Results. These information bits are designed to answer all the "five W's and two H's" questions (why, what, where, when, who, how, and how much), as well as to describe other aspects of the monitoring effort. Having the information that describe the data is essential when you need to make comparisons between data sets, because that information will tell you if you are looking at 'apples' or at 'oranges'. The NWQMC has established the Methods and Data Comparability Board (MDCB) to facilitate data comparability among all monitoring entities in the Nation, and the one of the Board's workgroups has developed a list of water quality data elements (WQDE) that are deemed essential for data interpretation and comparability. The list was adopted in 2001, and monitoring Programs across the Nation are implementing it into their efforts.

5.0 Sources and Resources

This IP is an integral part of the Data Quality Management (DQM) System implemented by the Clean Water Team, the Citizen Monitoring Program of the California State Water Resources Control Board.

For an electronic copy, to find many more CWT guidance documents, or to find the contact information for the CWT Coordinators, visit our website at www.waterboards.ca.gov/nps/volunteer.html

If you wish to cite this IP in other texts you can use "CWT 2006" and reference it as follows: "Clean Water Team (CWT) 2006. Good data made tangible, DQM IP-8.2.3. in: The Clean Water Team Guidance Compendium for Watershed Monitoring and Assessment. Division of Water Quality, California State Water Resources Control Board (SWRCB), Sacramento, CA."

References and complementary Guidance materials

DQM-IP-8.2.4 - Representativeness of Environmental Monitoring Data
DQM-IP-8.2.5 - Sampling Design Principles
DQM-SOP-9.3.2.2 Instrument/batch-specific error calculation and data validation

SWAMP Field Methods Distance Learning Course, Common Element B (QA)

The Methods and Data Comparability Board's page on water quality data elements (WQDE) http://wi.water.usgs.gov/methods/tools/wqde/index.htm

U. S. Environmental Protection Agency (USEPA) 2000. Guidance for the Data Quality Objectives Process, EPA QA/G-4. USEPA publication EPA/600/R-96/055, Office of Environmental Information, Washington DC. August.

U. S. Environmental Protection Agency (USEPA) 2002. Guidance on Environmental Data Verification and Data Validation, EPA QA/G-8. USEPA publication EPA/240/R-02/004, Office of Environmental Information, Washington DC. November.