DO YOU KNOW WHAT'S DRIPPING OUT OF YOUR PIPE (OF ENVIRONMENTAL DATA, THAT IS)?

Sharon J. Mosset
Quality Management Consultant
Advanced Sciences Inc.
477 N. Shoup Ave.
Suite 207
Idaho Falls, Idaho 83402

ABSTRACT

A key goal for every environmental project manager is the production of **appropriate, sufficient data of known and acceptable quality**. Regardless of what type of data is involved--geological, chemical, radiological--of utmost concern to the end user is the reliability of that data for use in technical analysis and decision-making. For data to be defensible, the manager must be able to demonstrate that it was controlled throughout its entire life cycle of **conceptualization, generation, quality assessment, storage, technical utilization, and ultimate archival.**

To aid in understanding the many internal and external aspects of managing the quality of environmental data, a comparison is made between the environmental data flow to the flow of oil down the Alaskan pipeline: the EXXON quality control specialist must be concerned not only with controls over the activities at the end of the pipeline, but also with the integrity of the pipeline itself along the way, and what may have been introduced to or lost from the flow of raw oil along the way. **An effective data management program will ensure not only the integrity of the framework (pipeline) that the data (oil) resides and is transported in, but also the integrity of the collected data (oil) itself.**

The overall purpose of a data management system is to provide a comprehensive, logical framework in which all collected data are received, entered, processed, retrieved for use, manipulated, and ultimately archived. An effective data management program should address both the **hardware/software component** and the **resident data component** of the system. There are several specific quality control tools that can and should be employed during the life cycle of environmental data to provide assurance of the data's quality. These two components are discussed, the various quality control tools identified, and an outline proposed for a data management plan that will provide a comprehensive, yet practical data management program.

INTRODUCTION

Regardless of what type of data is involved--geological, hydrological, radiological--of utmost concern to the end user is the **reliability** of that data, or how much risk exists in using those data in technical analysis or decision-making. A key objective for every environmental project manager is to produce appropriate data, of **sufficient quantity**, and with **known** and **acceptable quality**. For data to be defensible adequate control must be demonstrated throughout its life cycle of conceptualization, generation, quality assessment, technical evaluation, and ultimate archival.

Some general challenges to achieving the goals of high quality, defensible data include: 1) the many inherent risks associated with environmental investigations/assessments; 2) the lack of knowledge of the many types of data quality controls and the programmatic points that would most benefit from their application; and 3) the current absence of general industry guidance for developing a comprehensive, yet practical and cost effective, data management program.

To better grasp and appreciate the many external and internal aspects of managing the quality of environmental data, compare the environmental data flow to the flow of oil down the Alaskan Pipeline: the EXXON quality control technician must be concerned not only with controls over the activities at the end of the pipeline, but also with what was originally collected and the integrity of the pipeline itself (what may have been introduced to or lost from the flow of raw oil along the way). An effective data management program will ensure not only the integrity of the framework (pipeline) that the data (oil) resides and is transported in, but also the integrity of the collected data (oil) itself, from the collection point to its ultimate use.

The overall purpose of a data management system is to provide a comprehensive, logical framework (where framework is defined as the hardware/software and all programmatic documentation) in which all data generated are received, entered, processed, stored, manipulated, retrieved for use, and ultimately archived. These data activities must occur in a consistent, controlled, and clearly-traceable (documented) manner.

The data management system must perform its critical functions in a consistent, cost-effective, and defensible manner, while also addressing all applicable standards and guidance. In addition to remaining responsive to the dynamic regulatory climate and withstanding the scrutiny of various regulatory bodies and the public, this program must also offer functional flexibility. That is, while retaining the integrity of all components of the data management system and its resident data, the system should also allow for possible modifications to database software and/or hardware, e.g., converting from a personal computer-based system to a mainframe system.

This paper: 1) identifies the many types of data quality controls that exist, and the points along the pipeline (data flow) that the project manager can and should build in these controls; and 2) proposes an implementation strategy for building a defensible, practical data management program that will yield environmental data of known, acceptable quality.

DESIGN CONSIDERATIONS

The design for a data management system should consist of two key components: the hardware/software (framework) component and the resident data component. To help ensure that sufficient quality management elements are incorporated into these two key components, it is helpful to recall the close analogy between the flow of oil down a pipeline and the generation and use of environmental data. In both cases, it is extremely unrealistic to install quality control devices only at the end of the pipeline (or after the data already resides in a database) and expect to have confidence in the quality of raw oil (data) coming out of the pipe. Just as the EXXON pipeline quality control technician must be concerned with the oil from the point it is pumped from the ground (or ocean) until it is ready for shipment to the processing facility, the environmental project manager must consider all points in the data's life cycle, as well as the hardware/software framework in which it resides.

Hardware/Software Component

Some excellent guidance on software control can be found in Part 2.7 of ASME-NQA-2-1989, "Quality Assurance Requirements for Nuclear Facilities."(1) Part 2.7 supplements the requirements contained in the basic element of Design Control of ASME-NQA-1, and discusses the controls to be placed on software during each of its life cycle phases of:

- conceptualization,
- design,
- implementation,
- testing,
- verification and validation,
- installation and checkout,
- operation and maintenance, and
- retirement.

These Part 2.7 requirements apply to any "computer software used to produce or manipulate data which is used directly in the design, analysis, and operation of structures, systems, and components."(1) Whether the software is developed in-house or by a contractor, or purchased "off the shelf," adequate controls must be identified and documented in a general quality assurance plan, a software control plan, or as a component of a data management plan. Controls are particularly critical when the software are to be used for data manipulation or applications such as modelling for risk analysis or performance assessments.

Although written to support the collection of scientific and technical information for site characterization of a high-level nuclear waste repository (Yucca Flats), ASME-NQA-3-1989(2) is generally recognized to provide pertinent guidance for any QA program that is supporting scientific investigations (*field sampling*, sample and data handling, and data interpretation). The Design Control element (element three) contains a sub-element (3.5, Computer Software Control) that parallels the guidance found in Part 2.7, only translating the terms into environmental investigations and remediation, rather than the operation of nuclear facilities.

Once a particular configuration of hardware and software has been sufficiently tested and is operating successfully, it should be placed under formal configuration control. Procedures should be written for such functional data management activities as: data entry, data verification and validation, uploads, downloads, extractions, transfers, reduction, manipulation, storage and archival. One example of the level of detail that should be considered in these procedures, is the identification of the required form and format that field and laboratory organizations must use to submit data for entry into the master database, if prior verification is required, and how it must be performed and documented.

Resident Data Component

Again, referring back to the analogy with the oil pipeline, all phases of the environmental data life cycle must be considered when installing data quality controls; e.g., conceptualization, generation, quality assessment, technical utilization, and ultimate archival.

Conceptualization: The data life cycle begins with the identification of the need for a particular type of data. Historical data should first be evaluated to determine whether new data must be collected, and if so, what type, how much, and at what location. This evaluation of existing data introduces the first quality control tool: **data qualification**. All data that was not collected under the control of the project's current QA program must first be qualified for its adequacy in meeting the current intended use. A procedure should be developed to guide and control this data qualification process. This procedure can be attached to the work plan or the general quality assurance program plan (QAPP). The extent of qualification to be performed should be graded, based on the ultimate use for the data. An excellent reference for the data qualification process is NUREG-1298, "Qualification of Existing Data for High-Level Nuclear Waste Repositories"(3).

If no data currently exists that provides the information needed for a particular investigation, new data must be collected. To ensure that adequate planning is conducted (to establish the requisite level of quality) and to provide a means of evaluating whether or not the quality objectives have indeed been achieved, the EPA requires the development and application of **data quality objectives** (DQOs). DQOs are "quantitative and qualitative statements which specify the quality of data required to support Agency decisions during remedial response activities"(4). Quality objectives must be assigned for the five quality parameters of precision, accuracy, representativeness, completeness, and comparability. In simple terms, the use of DQOs provides the project manager with the means to determine the appropriate type of data, the best method by which to collect it, where to collect it, how much is needed, and how accurate it must be.

During the conceptualization phase, the two key quality control tools to consider and apply are <u>data qualification</u> and the <u>identification of DQOs</u> in plans and procedures.

Data Generation: The DQOs established during the data conceptualization phase should be referenced in the specific site investigation plan, and detailed in the associated site-specific QA project plan (QAPjP) and sampling and analysis plan (SAP). The specific means by which these DQOs are to be met should be addressed in the technical (field and laboratory) procedures. The field manager is responsible for achieving the DQOs established for representativeness, completeness, and comparability, and the laboratory must meet the DQOs set for accuracy and precision.

During their document review, the QA organization should verify that adequate DQOs have been established upfront in the appropriate planning documents. Once the site investigation is underway the QA organizations (field and laboratory) should perform surveillances and audits to verify that the procedures referenced in the QAPjP and SAP are being performed correctly, and by adequately and appropriately trained personnel.

During data generation, the detailed DQOs should be applied and the QA organization should verify that these DQOs have been established and implemented through document reviews, surveillances, and audits.

Quality Assessment: Once the data collection/analysis process has been completed, both field and analytical results must be verified and validated prior to their use by the technical project team. The EPA requires that: "specific routine procedures to assess precision, accuracy, and completeness on the project be described in each QAPjP; the principal criteria that will be used to validate data integrity during collection and reporting of data be established; and the methods used to identify and treat outliers resulting from assessment process be addressed(5)."

The quality assessment process includes both **data verification** and data validation activities. Data verification must be performed before data validation can begin. Verification is an **ongoing**, routine activity conducted by technical, laboratory, and administrative personnel on **small** sets of data to determine: if the data have been accurately quantified, recorded, and transcribed; collected and analyzed in accordance with the prescribed, current procedures; if the data appear suitably complete; and if the data appear to be reasonable and consistent, based on prior knowledge of the site being investigated.

Data verification should be performed on field records (logbooks, data collection forms, labels, chain of custody forms, shipping forms, requests for analysis forms) and all lab records (logbooks, quality control tests, lab results). Data entry verification should also be performed to compare hard copy source documents and database print-outs.

Validation can be performed on data that has been successfully verified. "Data validation is a systematic process for reviewing a body of data against a set of criteria to provide assurance that the data are adequate for their intended use. Data validation consists of data editing, screening, checking, auditing, verifying, certifying, and review."(5) Its ultimate purpose is to determine if the pre-established DQOs have been achieved. Symbols (called qualifiers or flags) are assigned to clearly indicate whether, in general terms, the data are acceptable as is, the values should be accepted as estimates only, or the data should be rejected.

As with the qualification of existing data, the extent of quality assessment to be performed should be graded, based on the intended use of the data and what level of quality (confidence) must be present to assure its usefulness. The quality assessment process should be complete and carefully documented before the data package is submitted to the project team for analysis and use in making decisions.

The quality control tools of data verification and data validation are applied during the quality assessment phase.

Technical utilization: The "technical evaluation begins on the date that the first set of validated data are received...and ends during the preparation of the report...includes reduction, tabulation, manipulation, interpretation, and environmental fate and transport modeling/evaluation."(6) The project manager should ensure that any data assigned qualifiers or flags are clearly identified, and explanations/justifications provided for the symbols assigned.

During the technical evaluation process, the project team should look for the assignment of any data qualification flags before utilizing the data for analysis, manipulation, decision-making, and report preparation.

Storage and archival: Data are usually entered into the master environmental database in their "raw" form by the organizations conducting field work or performing analysis. These data reside in temporary storage until they have undergone the data quality assessment process. As it is critical that the database provide for the secure storage, retrieval, uploading, downloading, transfer, and manipulation of validated data only, those data that have been assigned qualifiers must be clearly flagged in the system. Data that have been rejected as a result of the validation process must be removed from the database, or clearly flagged as invalid. Once the database has been updated, it should be "secured," and the data "archived," so that changes cannot be made without authorization by the database manager. The status of the residing data--preliminary or final-- must be clear at all times to all database users.

Data residing in the database must be clearly identified as validated or unvalidated. Data qualification flags must be entered and secured so that the technical users can utilize the archived data with confidence in their technical evaluations and decision-making.

IMPLEMENTATION STRATEGY

A top-tier data management plan (DMP) should be developed that addresses both of the above-described design components, as well as all the specific quality control elements that ensure the integrity of these two components. The DMP should reference (and be referenced by) the project's work plan, QAPP, any lower-tier QAPjPs for field work and laboratory analysis, and SAPs. The design is implemented through technical and administrative procedures. The DMP should contain the following elements, as a minimum:

- the identification of all organization(s) involved in data generation, entry, validation, manipulation and usage; and the specification of roles, responsibilities, lines of authority, communication, and interfaces required for developing, executing, and assessing the data management program;
- the types of data the program is to manage the data flow;
- the types of software to be used in managing the data;
- the specification of procedures by which the system is to be developed, tested, operated, controlled, modified, and documented;
- an assessment program that monitors DMP compliance and evaluates performance, and that ensures that problems are promptly reported, documented, evaluated, and resolved; and
- provisions for system security.

Detailed Implementation Strategy

To implement an effective, comprehensive, yet practical data management program, the following general steps should be followed:

- Identify all applicable general program requirements (federal, state, site, project)
- Research and evaluate specific data management program needs
- Draft data management plan (suggested content/format immediately follows text of paper); plan must tie in closely with other programmatic plans and procedures
- Develop implementing procedures
- Develop and conduct training program
- Develop and administer an assessment program; modify DMP and/or procedures as necessary

RECOMMENDED FORMAT/CONTENT FOR DATA MANAGEMENT PLAN

1. INTRODUCTION
    - 1.1 Purpose, Scope, and Applicability
    - 1.2 Requirements and References

2.0 ORGANIZATIONS, ROLES, AND RESPONSIBILITIES
- 2.1 Technical Contractors
- 2.2 Laboratories
- 2.3 Other Technical Support
- 2.4 Data Management Group
- 2.5 External Data Users (EPA, State, public)

3.0 DATA MANAGEMENT CONTROLS
- 3.1 Introduction - Types of Data and Data Flow
- 3.2 Technical (Field and Laboratory) Responsibilities *
    - 3.2.1 Establishing Data Quality Objectives
    - 3.2.2 Qualification of Existing Data
    - 3.2.3 Calibration Procedures and Frequency
    - 3.2.4 Preventive Maintenance
    - 3.2.5 Sampling Procedures
    - 3.2.6 Sample Custody
    - 3.2.7 Analytical Procedures
    - 3.2.8 Internal Quality Control Checks
    - 3.2.9 Performance and System Audits
    - 3.2.10 Specific Routine Procedures Used to Assess Data Precision, Accuracy, and Completeness
    - 3.2.11 Data Reduction, Validation, and Reporting
- 3.3 Data Management Group Responsibilities
    - 3.3.1 Data Entry
    - 3.3.2 Data Verification (New and Existing Data)
    - 3.3.3 Data Retrieval
    - 3.3.4 Data Uploads, Downloads, Transfer
    - 3.3.5 Data Manipulation
    - 3.3.6 Data Storage and Archival

4.0 SOFTWARE MANAGEMENT CONTROLS **
- 4.1 Introduction - Software Life Cycle and Types of Software Used
- 4.2 Software Procurement
- 4.3 Software Development
- 4.4 Software Testing
- 4.5 Configuration Control

5.0 ADMINISTRATIVE CONTROLS
- 5.1 Training
- 5.2 System Security
- 5.3 Problem Reporting and Corrective Action
- 5.4 Records
- 5.5 Assessment

Appendices
- A - Glossary of Acronyms and Definitions
- B - Listing of Data Management Procedures (title, scope, and schedule for development
- C - Listing of Software Management Procedures (title, scope, and schedule for development)

* Elements based on QAMS-005/80 requirements

** Elements based on NQA-2, Part 2.7 requirements

REFERENCES

1. American Society of Mechanical Engineers (ASME)/NQA-2a-1990 Edition, Quality Assurance Requirements for Computer Software for Nuclear Facility Applications.
2. American Society of Mechanical Engineers (ASME)/NQA-3-1989 Edition, Quality Assurance Program Requirements for the Collection of Scientific and Technical Information for Site Characterization of High-Level Nuclear Waste Repositories.
3. Nuclear Regulatory Commission (NRC) NUREG-1298, 1988, Qualification of Existing Data for High-Level Nuclear Waste Repositories.
4. U.S. Environmental Protection Agency (EPA/540/G-87-003, March 1987), Data Quality Objectives for Remedial Response Activities: Development Process.
5. U.S. Environmental Protection Agency (QAMS-005/80), February 1983, Interim Guidelines and Specifications for Preparing Quality Assurance Project Plans.
6. U.S. Environmental Protection Agency (EPA/540/G-89/004, October 1988) Guidance for Conducting a Remedial Investigation/Feasibility Study Under CERCLA.