Environmental Data Management Best Practices

Data Quality Dimensions
Applicability and Data Usability for Environmental Data Management

Data quality is a broad topic that encompasses multiple different considerations, or dimensions, each of which may be of varying importance to different activities that are conducted throughout a project's lifecycle. Managing data quality can be simplified by addressing each of these dimensions individually. This document defines five principal data quality dimensions and provides examples of data quality items in each dimension that apply to lifecycle stages, such as field data collection, historical data collection, and data reporting. This document also addresses the relationship between data quality and data usability

Overview

Data quality can be easily understood as an overall concept, but it encompasses many potential considerations and practices that can be difficult to summarize both comprehensively and succinctly. A comprehensive knowledge or review of relevant considerations and practices is necessary to implement a practical program to assure, or to assess, the quality of environmental data. This document describes "data quality dimensions" as a tool for both thinking about and organizing the considerations and practices relevant to each program or project.

Environmental data are frequently generated by processes that provide a broad scope for errors resulting from faulty recording, transcription, encoding, tabulation, presentation, and documentation. These types of errors result in data values that are not correct. Poor data quality may also result from omissions; for example, sampling depths or depth units weren't recorded, coordinates are not included in a data table because they are in a shapefile, and summary tables omit information so that they fit on the page better. These types of omissions result in a data set that is not complete. Completeness and correctness are examples of two data quality dimensions. A data set may be complete, yet contain incorrect values, and conversely, all the values in a data set may be correct, but the data set (as a whole) may be incomplete. These and other data quality dimensions (integrity, unambiguity, and consistency) are defined and described more fully below in the Dimensions of Data Quality section of this document.

The applicability and importance of each data quality dimension can vary for different stages of the data lifecycle, and for different data uses. For example, there may be many completeness considerations to address during the stage of field sampling, but fewer and different considerations to address during the stage of data reporting. The use of a data set for risk assessment will require a high level of data quality in the correctness dimension, whereas the use of a data set for planning of future sampling may have lower requirements for correctness. Data applicability and data usability are described more fully below under the Applicability of Data Quality Dimensions and Assessing Data Quality sections of this document.

Importance of Documentation

When data quality, applicability, and usability are established or assessed, the results must be recorded and made available to all data users. Documentation is therefore an essential component of data quality. Lack of documentation of a data set's quality compromises the usability of those data.

There are several types of documentation that are important to support the establishment, verification, or assessment of data quality. These include:

- Planning documents. This includes work plans, quality assurance project plans (QAPPs), sampling and analysis plans (SAPs), and data management plans (DMPs). This may also include a customized data quality implementation matrix (see below under the Applicability of Data Quality Dimensions) and related information.
- Provenance, history, and related metadata. This includes descriptions of the origin, content, purpose, rights, and revision history of each data set.
- Implementation documentation. This consists of documentation of the fulfillment of the requirements specified in the planning documents, and of any deviations from those plans. This includes field records, laboratory reports, data reports, and other documentation that describes methods and other details pertaining to sample collection and data generation.
- Operational process documentation. This includes descriptions of how data were handled during acquisition, cleaning, standardization, and summarization. This also includes data migration plans and processes; data audits; and data quality assessments and processes used to establish data quality, such as data corrections. This should also include records of data changes that may be made after initial data acquisition. Some of this information may be present in a prospective form in a DMP, but data sets may include features that require special handling rules or processes that are not described in the DMP, but that are essential for proper data integration and interpretation.
- Data usability assessments. For each potential use of a data set (for example, site characterization, risk assessment), the usability of the data set must be assessed and documented. Data usability assessments may be based on the results of data quality assessment (see the Dimensions of Data Quality and Assessing Data Quality sections below).

Field sampling activities use well-established chain-of-custody (COC) procedures to ensure the integrity of environmental samples from collection through laboratory analyses, but there are no comparable standard chain-of-custody processes for data following their generation by laboratories, or data that are obtained from other sources. Documentation of data provenance and handling should be sufficiently complete and detailed that it can serve the same purpose: that is, to ensure that every data value is traceable from its origin through to its ultimate use.

In addition to supporting traceability, documentation of operational processes allows quality assurance checks of the methods used for data management. The operational documentation should be sufficiently detailed that it allows those processes to be repeated exactly, yielding the same outputs from the same inputs. This documentation may take several forms, including narrative descriptions, tables, and (well-annotated) program code, such as SQL scripts. Operational processes may change over the duration of a program or project, and the documentation of those processes should be kept up to date. The history of changes to the documentation should be preserved by using a version control system or separate sequential backups.

Documentation of data quality assessments and resolution of data quality problems is particularly important when the same data set may be used by multiple organizations, and there is the potential for discrepancies between versions of the same data set that have and have not had data quality issues rectified.

Procedures for creation of adequate documentation should be incorporated into an organization's standard data management practice so that they are applied reliably and consistently. Best practices for data management should be established, and should specify the level of documentation that is needed by the organization or for each project or data set. This documentation should be created in a form, and collected in a location, that is accessible to all data managers, data users, and stakeholders.

Dimensions of Data Quality

Potential data quality issues can be categorized into several fairly distinct data dimensions, each of which addresses a specific aspect of data quality. The five data quality dimensions described in this document are:

- Data integrity
- Unambiguity
- Consistency
- Completeness
- Correctness

Each data quality dimension is described in one of the following sections. Each of these dimensions can contain multiple elements. For example, in a field sampling program, completeness applies to the set of locations sampled, the number of samples collected, and the number of analytical results returned by a laboratory. These data quality dimensions represent distinct aspects of data quality, but they are also interrelated (for example, data integrity issues may result from a lack of completeness in a data set).

The level of data quality for each element on each of these dimensions may be represented as a scalar or categorical quantity (that is, it can have a value ranging from low to high, or bad to good). Although ideally every data set is at the top of the scale on all five dimensions, this is not always the case in practice. Cost and feasibility, among other things, may limit a program's ability to achieve the highest level of data quality in all respects. Prior to collecting and compiling data, it is therefore important to establish the level of quality needed for the task being implemented. Trade-offs may sometimes have to be made between data quality elements and other program goals. Any such trade-offs should be identified during project planning and incorporated into the project's data quality objectives (DQOs).

Data Integrity

Data integrity means that different types of information are properly identified and related to one another. For example, every client sample identifier contained in a laboratory deliverable should match a unique sample identifier from field records. There are three elements of data integrity:

- Uniqueness—For example, every sample has a unique identifier.
- Relational integrity—For example, client sample identifiers used by laboratories all correspond to samples collected by the field program.
- Domain integrity—For example, concentration measurements are all represented by numerical values.

Data integrity problems are common in hand-entered data and are also often found when different parts of a data set are received from different sources, such as field records and laboratory records. Data integrity problems may be introduced when information is not recorded either through error or because, for example, field forms were not designed to accommodate it. Data integrity problems can also be introduced when inconsistent or incomplete updates are made to data stored in forms such as spreadsheets. Data integrity problems can also be introduced during data exchange as a result of data truncation, differences in data models, and differences in domains of valid values or mistranslation of valid values.

Relational databases can be built with rules that enforce data integrity, and particularly in large data sets, data integrity failures may be difficult to find until such constraints are applied to the data. Depending on the type of relational database management system used (see the Environmental Data Management Systems White Paper), these rules can include primary keys, other unique keys, foreign keys, check constraints, and trigger procedures.

Data entry software can be designed to check and prevent data integrity errors, and can provide more informative error reporting than might be produced by the underlying database management system (DBMS) or other data storage format.

Unambiguity

Unambiguity means that table column names and data values (such as sample identifiers and analyte codes) have a single and unique meaning. Ambiguity occurs when an item may have multiple meanings or when two or more items appear to have the same or very similar meanings. Examples of ambiguities that are found in environmental data sets include:

- Column names. Is a column named “Sample type” meant to distinguish between sediment, water, and tissue samples; between natural and field quality control (QC) samples; between site and background samples; between original and confirmation samples; between bench and pilot test samples; or something else? Column names such as “Depth” may appear in several tables, representing different quantities (for example, water depth and corer penetration depth), and their meaning may be clear in those contexts, but become ambiguous when the data tables are joined in a query.
- Coordinates. Geographic coordinates that are not accompanied by a spatial reference system identifier (that is, a datum and possibly a projection) are ambiguous. For example, coordinates in decimal degrees may be based on either a NAD83 or WGS84 datum, and the difference may amount to a perceptible and important difference in where locations are shown on maps.
- Data attributes that are embedded in sample identifiers. Sample identifiers frequently contain sections that identify the location, sampling phase, material collected, and other information (for example, “MW1-1025-GR”). The values used in these fields are often abbreviated and non-obvious, and their meanings may be undocumented. Typographical errors in such identifiers can be hard to recognize and can introduce further ambiguity.
- Valid values. For example, a data set may contain analyte codes of “PAH” and “TPAH” without a clear distinction between their meanings. Analyte codes for hydrocarbon ranges can also be ambiguous if the minimum and maximum lengths of the carbon chain are not explicitly specified.

Consistency

Consistency means that the same type of information is always represented in the same way. Consistency issues frequently arise when integrating data from different sources. Valid value lists ordinarily differ between different data sources. Inconsistency in the use of valid values is sometimes also found within individual data sets. Consistency issues often underlie issues of ambiguity and completeness.

Data sets may differ in the consistency of data structure and reporting detail. Example 1: One data set may include analytical results for individual laboratory replicates, whereas another may contain only average results for each location and date. Example 2: One data set may contain a detailed description of the type of material collected and analyzed, whereas another may contain only the laboratory’s characterization of the material (so that sediment is identified as soil, for example).

Completeness

There are two aspects of data completeness:

- All of the relevant data sets are in hand.
- Each data set contains all the required data.

Completeness of an individual data set may be assessed with regard to minimal data requirements (for example, location, date, and depth information is all present) and with regard to intended uses of the data (for example, chemical analyses and toxicity tests are conducted on splits of the same sample). A data set may be complete in that it satisfies minimal requirements but is nevertheless incomplete for the purpose of conducting a particular analysis.

Correctness

Correctness for environmental data means that the measurements are accurate and sufficiently precise in terms of representing environmental conditions at the location, date, and depth indicated. Correctness of analytical chemistry data is ordinarily assessed by data validation. There are no equivalent rigorous and standardized procedures for other types of environmental data, so processes for verifying correctness of other types of data may need to be developed and applied for each program or project.

Various translations and transformations may be carried out to standardize data (such as converting units) and resolve data quality issues. These processes must be carried out in a way that maintains data correctness and fidelity to the original source. Some deviations from the original data source may, however, legitimately result from corrections or clarifications to resolve data quality issues.

Applicability of Data Quality Dimensions

The five data quality dimensions provide a useful framework for classifying different types of data quality issues, but there are additional considerations relevant to their actual application. One of these is the different stages in the data lifecycle. For programs or projects that are engaged in the collection and management of environmental data, these lifecycle stages may include the following:

- Field work, including sample collection and recording of other measurements and observations
- Laboratory analyses and data validation
- Acquisition of data from historical sources
- Acquisition of geospatial data
- Acquisition of unstructured data (photographs, videos, printed media)
- Acquisition of generational knowledge
- Synthesis of data from multiple sources to create a comprehensive and consistent repository
- Data summarization and reporting

These lifecycle stages fit into the more general steps described in the Data Lifecycle Fact Sheet.

Specific data quality issues will ordinarily differ not only by data quality dimension, but also by data lifecycle stage. Examples of quality issues that are specific to each data quality dimension and lifecycle stage are shown in Table 1.

Table 1. Application of Data Quality Dimensions to Project Lifecycle Stages

Project Stage	Activity	A. Data Integrity	B. Unambiguity	C. Consistency	D. Completeness	E. Correctness	F. Documentation
Acquire	**I. Field work, including sample collection and recording of other measurements and observations**	1. The location is recorded for every sample collection	1. Location identifiers are consistent with any previous use	1. Location, sample, and analysis identifiers follow the planned design	1. All planned locations are visited	1. All SOPs are followed	1. Deviations from the sampling plan are fully described
		2. The date is recorded for every sample collection	2. Spatial reference system used is	2. The same date format is used throughout	2. All planned natural samples are collected	2. All manual entries in logs and forms are checked	2. Field sheets and log pages are signed
		3. Information is recorded to clearly define the sample location/type	3. Dates and times are complete	3. Capitalization and use of dashes and underscores are	3. All planned field blank or other QC samples are collected	3. Sample labels match field logs and forms	
		4. Sample attributes are recorded as appropriate to the material, location, and other sampling conditions	4. Units for depths and other field measurements are recorded			4. The correct analyses are designated for each sample	
		5. Field replicates and splits are distinguished	5. Any abbreviations used are documented in planning documents or			5. Sample labels match COC and analysis request documents	
		6. Field blanks are distinguished from natural samples	6. Sampler names are recorded in full			6. Correct COCs and analysis request documents are shipped with samples	
		7. The relationship between collections and subsamples is recorded					
	II. Laboratory analysis and data validation	1. Relationships between lab and client sample identifiers are recorded	1. All abbreviations and codes used are defined	1. Client sample identifiers used to log samples in at the laboratory match those	1. All requested analyses are completed	1. All laboratory SOPs and method procedures are followed	1. Deviations from standard reporting formats or content are documented
		2. Reanalyses and replicate analyses are distinguished from original	2. All laboratory flags and qualifiers used are defined	2. Approved and appropriate analytical methods used.			
	III. Historical data acquisition	1. Unique identifiers are used for each data set (historical and current).	1. Codes used in all data sets are defined	1. Multiple sources of the same data have consistent results	1. Methods to find relevant data have been reviewed by knowledgeable	1. Data extracted from PDF files or other data sources is double-checked	1. An inventory or list of historical data sources is maintained
			2. Conflicting location identifiers and coordinates are resolved.	2. Valid values are standardized or added to eliminate conflicting definitions from different sources.	2. Work plans and field sampling plans for each data set have been obtained		2. Data gaps and inconsistencies are recorded.

Project Stage	Activity	A. Data Integrity	B. Unambiguity	C. Consistency	D. Completeness	E. Correctness	F. Documentation
Acquire	IV. Spatial data acquisition	1. Each location identifier is associated with only a single location	1. The spatial coordinate system, datum, and units of measure are clearly identified (e.g., survey foot vs. international foot)	1. The organization's map publishing standards and cartographic best practices are followed	1. All required metadata is recorded for every geospatial measurement and data set	1. Accuracy and precision are documented and maintained	1. Complete metadata is recorded for each data set collected or used, per organizational or contractual requirements
		2. Coordinates are always accompanied by a coordinate reference system identifier	2. A scale bar and north arrow are clear on all maps	2. Map scales are similar or identical for all data layers used for analysis and presentation	2. Geospatial data is collected for every required sample or measurement	2. Geolocation uncertainty is documented	2. Map publication standards and cartographic best practices are documented
			3. Contour intervals are clearly identified			3. Field data post-processing rules are followed	3. A narrative revision history is maintained for data sets that are changed or replaced
			4. Data sources for maps are identified			4. The spatial coordinate system, datum, and units of measure are clearly identified (e.g., survey foot vs.	4. Data collection files are preserved by archiving or backup
			5. Map legends show all symbology used			5. Distances and areas are calculated in a rectangular (projected) coordinate	
	V. Traditional knowledge (generational knowledge) data acquisition	Ensure that the data maintains the data context.	1. Accurate language translation and cultural understanding.	1. Multiple sources of the same data have consistent results	1. Verify completness goals met	1. Accurate language translation and cultural understanding.	1. An inventory or list of data sources is maintained
			2. Data are from a trusted community source	2.Consistent keyword tagging		2. Data are from a trusted community source	2. Data gaps and inconsistencies are
						3. Consider weather or seasonal changes in relation	3. Special terminology or interpretations are
						4. Consider potential bias of data provider towards data collector and vice versa.	4. Special processes for acquisition and organization of TEK are documented.

Project Stage	Activity	A. Data Integrity	B. Unambiguity	C. Consistency	D. Completeness	E. Correctness	F. Documentation
Process/ Maintain	VI. Synthesis of data from multiple sources	1. Location, date, and depth information is available for every sample	1. Definitions are clear for every analyte code and other code	1. Codes are standardized for all data sets	1. All relevant data sets are available	1. Analytical chemistry data have been validated (if required) and qualifiers	1. The source, purpose, rights, and constraints of each data set are recorded
		2. 'Parent' samples have sample descriptions	2. Codes are not duplicative, conflicting, or overlapping	2. Data structure is standardized for all data sets	2. All data from each data set can be compiled together	2. Location coordinates are consistent with maps or are otherwise plausible	2. Data quality issues addressed during data synthesis are described
		3. Sampling information is available for every analytical result	3. Definitions are clear for every column name		3. Verify completeness goals are met		3. Data handling, cleaning, and standardization processes are described
			4. Geographic coordinate systems are specified				
			5. Units are specified for all measured values				
Publish/ Share & Retain	VII. Data summarization and reporting	1. Table joins are formed correctly	1. A data dictionary for table names and column headers is included	1. Table and column names, and code values, are consistent across multiple summaries	1. All requested components and contents of the summary have been produced	1. Data selection criteria have been correctly applied	1. The date and time that the summary was produced is provided
			2. Dataset-specific and global location identifiers are included if they differ			2. Data handling rules (e.g., treatment of nondetects) have been correctly applied	2. The name and date of data summarization scripts are provided
			3. The spatial coordinate system is identified			3. Numerical results are presented with an appropriate number of	3. Deviations from data handling standards are recorded
						4. Non-detects are presented as desired (for example, as detection limits or quantitation limits), and appropriately qualified.	4. A data dictionary for tables and columns is provided
						5. The summarization script and/or the summary has a quality assurance review before delivery.	5. Any usage restrictions are described
						6. The data summary is delivered to the correct individual(s).	6. Contacts are provided for requesting additional information or providing data corrections

The items contained in each cell of the matrix in Table 1 can be used as a checklist of data quality issues or considerations. For each of these checklists (that is, each cell in the matrix), the following questions should be asked:

- Are there any additional items that are relevant to the program or project?
- Are any of these items irrelevant or not applicable to the program or project?

For each of the applicable items, the following questions should be asked:

- What is the appropriate metric to quantify data quality for this item? These may be quantitative or categorical metrics.
- What methods or processes can be used to assess data quality for this element?
- How can this element of data quality be established if it is lacking?
- How can this element of data quality be maintained after it is established?

The answers to these questions should be documented in the project plan if they are addressed before work is initiated, or else in a data management plan or other operational documentation that is updated during the course of the project.

In addition to categorizing data quality issues by data quality dimension and lifecycle stage, it may also be useful to further categorize them by data type. The lifecycle stages shown in Table 1 do this to some extent, for example by distinguishing between analytical chemistry data and geospatial data. Depending on the nature of a specific program or project, additional distinctions between data types may be appropriate. This document does not include this additional level of categorization of data quality issues, but the value of including this detail should be considered by project leaders and data managers.

Assessing Data Quality

Both methods and timing should be considered when planning data quality assessments. Some methods are standardized, such as validation of analytical laboratory data, but assessment methods for other types of data may have to be developed and applied on an ad hoc basis, and may vary from one project or practitioner to another. Examples are the use of check plots for coordinate data and the screening of measurements for implausible magnitudes and units. When nonstandard methods are used, they should be comprehensively documented and the methods should themselves be subjected to a quality assurance review. Organizations may wish to develop their own standards for data quality review, and document them as standard operating procedures.

Data quality may be assessed at several different stages in the project lifecycle, and even at multiple stages. Assessing data quality as close to the point of origin as possible is recommended, because that is likely to provide greater opportunities for correcting the data. For example, during sample collection, review of all field forms, notes, COC forms, and containers at the end of every day, prior to sample shipping, increases the chances that errors and omissions can be corrected with less cost than if those problems were recognized only weeks later.

Synthesis of data from multiple sources is a lifecycle stage where data quality assessments ordinarily should be performed, because data from different sources are often inconsistent in their use of valid values and different in data structure and corresponding data integrity rules.

More rigorous or more specific data quality assessments may also be performed as part of data usability for specific analyses. If no data quality assessment is performed, data uses may be compromised by data quality issues of unknown type and severity.

The results of data quality assessments should be documented. Well-established standards exist for documenting the quality of analytical chemistry data (that is, data validation qualifiers) and geospatial data (that is, metadata), but not for other types of environmental data. Environmental characterization programs should therefore establish their own standard practices and formats for documenting data quality. Alternative forms of documentation may include purely narrative descriptions, tables of checks performed and their results, and categorical representations that are stored in the project database. In some cases, data quality may be most usefully characterized in terms of the uncertainty or variability of the data.

Establishing and Maintaining Data Quality

General guidelines for establishing and maintaining data quality are provided in the following sections, grouped by data quality dimension. These guidelines can be used as a basis for more specific and detailed procedures that are applicable to each of the elements in Table 1.

Resolving data quality issues often requires information and judgment from project staff in multiple roles: field staff, laboratory coordinators, data managers, data users, and project or program managers. The information and decisions used to address data quality issues, and the level of data quality that is established, should be documented. This may be done by incorporating data quality annotations or comments into the data set itself, by maintaining a log specifically for data quality issues, or by recording the information in operational process documentation.

Data Integrity

The goal of ensuring that data items, and groups of data items, are properly related to one another can often be most effectively met by structuring the data in a relational database. Relational databases can be implemented in such a way that they enforce data integrity constraints. Not-null constraints should be applied to each required table column, a primary key should be defined for each table, foreign keys should be established to enforce referential integrity, and column domains or check constraints should be used to enforce other required conditions.

Violations of data integrity can be subtle and require multiple strategies to detect and correct. For example, constraints may

be applied to a table column for concentration data to ensure that values are not-null and numeric. However, data may be received or entered that uses values such as “-999” to represent missing data; these values would not violate the not-null or data type constraints. Additional check constraints may be required to prevent the introduction of such “junk” data values. These check constraints may be part of the DBMS table definition (if the DBMS supports check constraints) or may have to be applied as part of the data entry or loading process.

Migration of a data set that is lacking data integrity into a database that enforces data integrity requires that the integrity issues be addressed. This may be done by:

- Acquiring additional information to resolve the problems. This information may come from work plans, field forms, reports, or interviews with individuals responsible for data collection or preparation.
- Making and documenting assumptions about which data sources or data items are most reliable. For example, if there are multiple collection dates for the same sample identifier, either the identifier or the recorded date may be assumed to be the correct value. Such assumptions may be based on overall characteristics of the data set, such as the rules used to construct sample identifiers or the consistency of a sequence of sampling dates.
- Making and documenting decisions about acceptable compromises to data quality. For example, if there are multiple collection dates recorded for a sample, if date is not important to data use, then a decision may be made to use either the first or last recorded date; if date is important, then more effort will need to be expended to resolve the issue.

Unambiguity

The approach used to resolve ambiguities varies depending on the nature of each ambiguity, but may include:

- Searching for clarifying information in work plans, final reports, and other related documentation
- Factoring codes into two or more distinct sets of unambiguous valid values
- Modifying column names to clarify their meaning
- Making (and documenting) an assumption about the meaning of an ambiguous value based on the presence of other similar or different values, or on the nature of associated data
- Omitting the ambiguous data if its meaning cannot be determined

Ambiguity often arises because of unstated context or assumptions that were known to the data originator, but not to later data users. Those data users may not recognize the ambiguity because of their own context or assumptions, which may not be the same as those of the data originator. If ambiguity cannot be resolved, adding data quality indicators into the data set is recommended, so that the information is readily available to data users.

Consistency

Approaches to resolving consistency issues include:

- Translating valid values to a common domain that is used by all data sets.
- Transforming coordinate data from one spatial reference system to another.
- Converting numeric values to common units.
- Modifying a database’s data model to accommodate distinct features of a new data set. (Note that inappropriate use of this action can lead to greater inconsistency between data sets.)

Although some general methods can be used to establish consistency (such as using translation tables for valid values), resolving data consistency issues often requires data correction or completion steps that are specific to each data set.

Completeness

When planning documents are available, completeness can be assessed by comparison of planned to actual results. Evaluation of the number of locations sampled, number of samples, and number of analytes measured per sample may point toward completeness issues even in the absence of planning documents. The preferred method to resolve completeness issues is to obtain the missing information from accessory sources of information such as work plans, field sampling plans, field logbooks and notes, laboratory data packages, and data reports.

When data are partially incomplete and supporting documents are unavailable or are themselves incomplete, a fallback approach is to use default or synthesized values. Examples of this approach are:

- When dates are provided only to the month, but dates must be specified to the day, the first day of the month is assigned.
- When location identifiers are missing, they are created from the sample identifiers, a geohash of the coordinates, or other available information.

Correctness

Correctness of analytical chemistry results is ordinarily assessed by data validation. Correctness of other types of information can be assessed by check plots of location coordinates, gaps or overlaps in sequences such as dates or depth intervals, use of unusual valid values, and outliers in the distribution of numeric values.

Correctness issues are often identified only when data are used for a specific analysis—that is, late in the project data lifecycle. Because it is possible that those data have already been used for some purpose, such corrections should be thoroughly documented. The documentation should include an explanation of the rationale for choosing to change pre-existing data. That documentation should be made available to, or transmitted to, individuals who have previously used the data or who may currently be using the data. Original data providers, if any, should also be informed of the data corrections.

Data Usability

The usability of a data set ordinarily depends on the intended use of those data, for which the relevance of different data quality dimensions may vary. Data usability should be considered during the planning process—for example, by establishing DQOs (USEPA 2006). However, after it has been collected, data are often used for purposes that were not envisioned at the time of collection. In such cases, a post hoc data usability assessment should be carried out. This can follow the same general approach as the DQO process, which is:

- Define the goals of the intended data use.
- Establish the methods to be used to analyze the data to meet those goals.
- Identify the type(s) of data to be used, and the data quality characteristics needed (for example, precision, accuracy, completeness).
- Evaluate the available data with respect to the characteristics needed.

The categorization of data quality items illustrated in Table 1 can be used to carry out the last of these steps, if those assessments have been completed and documented. In some cases, data usability may be best characterized in terms of uncertainty or variability.

Use of data for purposes outside the original goals for which the data were collected may require greater emphasis on some dimensions of data quality than were originally needed. Therefore, data usability assessments may lead to additional data quality assessments.

For additional information related to data usability and how it relates to data review, see the Analytical Data Quality Review: Verification, Validation, and Usability Fact Sheet.

Resources

- USEPA. 2006. Guidance on Systematic Planning Using the Data Quality Objectives Process, EPA QA/G-4. EPA/240/B-06/001. Office of Environmental Information, United States Environmental Protection Agency, Washington, D.C. February 2006.