2004

A Data Integration Framework to Support Triad Projects

Jim Mack
New Jersey Institute of Technology

Deana M. Crumbling
U.S. Environmental Protection Agency

Fred Ellerbusch
New Jersey Institute of Technology

A Data Integration Framework to Support Triad Projects

Jim Mack

Deana M. Crumbling

Fred Ellerbusch

Cost-effective and efficient site remediation and scientifically defensible decisions require site characterizations that are representative of site conditions. The Triad conceptual site model (CSM) is at the center of a continually improving site characterization process that begins during systematic planning and ends after the last data are developed. To gain the full benefit and greatest cost-effectiveness, the process of CSM refinement should be performed in real time. Thus, the use of collaborative data is critical for evolving and maturing the CSM. In the field, through the use of all available data that are of known quality, a skilled and experienced field team can collect sufficient site information to mature the CSM in a timely manner. To facilitate the planning and execution of such a process, an easily understandable framework is needed to structure data quality that supports scientifically defensible decisions and efficient projects. This article explores such a framework. © 2004 Wiley Periodicals, Inc.

INTRODUCTION

The full benefits of the Triad approach are realized when systematic planning to manage decision uncertainty is combined with dynamic work strategies (Robbat, 1997) and real-time measurement technologies throughout the project life-cycle of characterization, remediation, and site reuse (Exhibit 1; see also Crumbling et al., 2001; US EPA, 2001). Systematic planning is the period in the project when the conceptual site model (CSM) is developed. The CSM serves as a basis to reach agreement among the various stakeholders. It is also the time when project endpoints are clearly articulated, the range of remedial approaches defined, clean up criteria established, and the general investigation methods and procedures developed.

Dynamic work strategies are integrated systems of decision logic and rules that define how field decision making will proceed using the information provided by the real-time tools selected for the project. The decision logic and rules are linked to the CSM, such that as information on site contaminants and conditions are developed through real-time data collection, the CSM is tested and continuously refined to account for newly discovered conditions. The decision logic and rules are developed by consensus among the project team and stakeholders. The written decision logic guides the field team along approved lines of decision making while enabling the investigation to proceed efficiently in real time.

Real-time measurement technologies include all the tools that generate and manage site information rapidly enough to support dynamic work strategies. They include a broad range of available tools for contaminant sampling and analysis, soil and geo-

Published online in Wiley Interscience (www.interscience.wiley.com). DOI: 10.1002/rem.20030

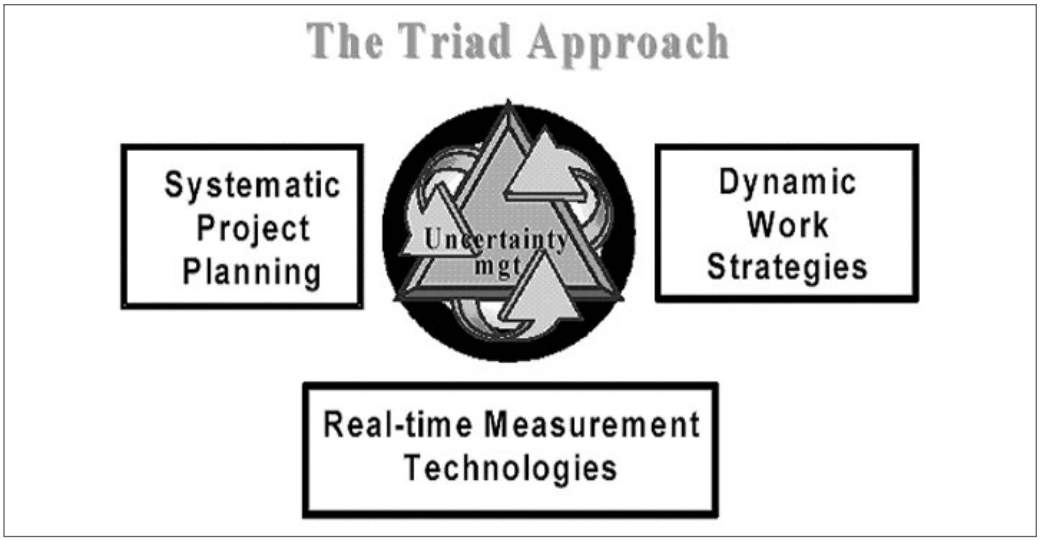

Exhibit 1. The Triad approach components

physical characterization, and location information (geographic position system), as well as data management and display software. Any data generated in the field are useable as long they are of known quality, appropriate to the decision at hand, and based on prior stakeholder agreement.

Literature on the use of innovative site management strategies, including the Triad approach, is emerging in the form of research, government and standards organization guidance, and case studies (for more information about the Triad approach, see American Society for Testing and Materials, 1998a, b, 1999; Applegate & Fitton, 1997; Burton et al., 1995; Connecticut Department of Environmental Protection, Leaking Underground Storage Tank Program, 2000; Crumbling et al., 2001; Ellerbusch et al., 2004; Mack et al., 2003; Robbat, 1994, 1997; Texas Natural Resource Conservation Commission, 1995; US EPA, 1997, 2000, 2001; Woll et al., 2003).

To assist with the level of data produced in the field through a Triad approach investigation, a predefined data management system is needed to store and use data in real time and when overlapping collaborative methods and procedures are used to manage both sampling and analytical aspects of data uncertainty. It is also useful for data quality control.

Data quality is controlled through two primary mechanisms:

1) reviewing real-time results to evaluate whether the sampling and analytical techniques are performing as expected (i.e., are in control); and
2) comparing real-time new data with the evolving CSM to detect incompatibility.

Iterative reciprocal real-time compatibility checks between the CSM and the sampling and analytical findings (data) serve as a powerful quality check that is only available in dynamic work strategies. A finding of incompatibility between the data and the CSM triggers investigative processes to determine whether problems have arisen in the sampling/analytical process (so they can quickly be corrected) or whether the existing CSM is inaccurate (requiring a revision of the CSM to match the new information).

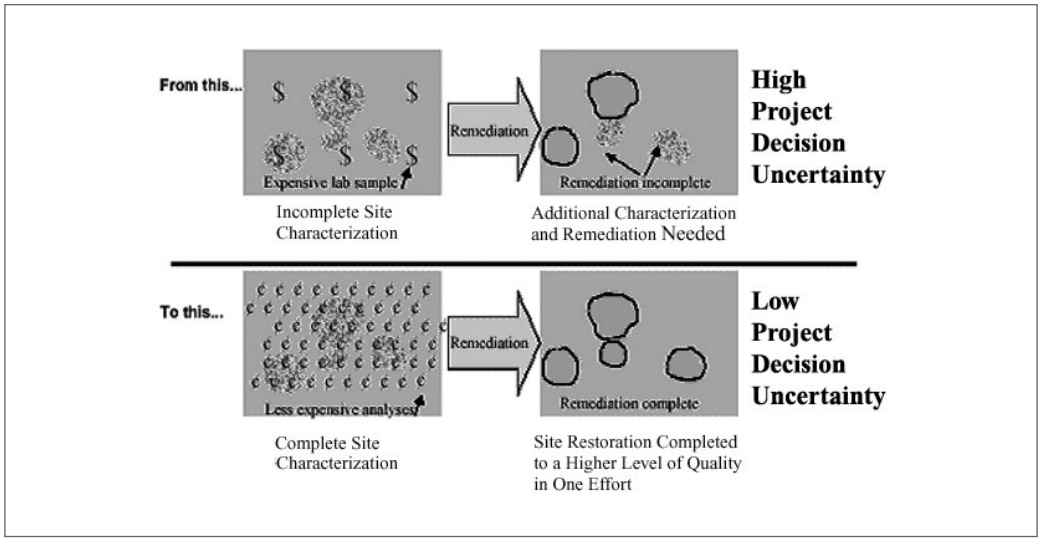

Exhibit 2. Sample representativeness and uncertainty; by collecting a larger number of less expensive (¢) samples, a more complete understanding of site conditions can be achieved

MATRIX HETEROGENEITY AND DEVELOPMENT OF ACCURATE CSMS

Environmental matrices are heterogeneous at both the macro and micro scales. Hence, generating representative data to account for the spatial heterogeneity within environmental matrices is critical to develop an accurate CSM. Although sample data may be correct in the sense that the analytical results are accurate for the tiny amount of material analyzed, extrapolating the results to a much larger volume encompassed by subunits of a site, or the entire site, may create a false picture. This is considered a form of sampling error. Sampling error can contribute to misleading CSMs even though the analytical method is performing correctly, possibly leading to erroneous decisions about risk or remediation strategies.

Under the traditional approach to site characterization, with its reliance on an Area of Concern (AOC) approach and high-quality but expensive analysis, usually too few samples are collected to develop an accurate picture of site conditions. There are a multitude of underlying reasons and reinforcing factors that create a false picture, including:

- AOCs are, by definition, biased toward contamination.
- Representative sitewide sampling is generally counterintuitive to AOC-focused sampling.
- More expensive high-quality analytical techniques reduce the number of samples in a given sampling budget.
- Regulation-based characterization regimes often prescribe in detail the approach to be used, often emanating from an AOC-driven preliminary assessment.

When too few samples are collected, there is little choice but to extrapolate the results of a tiny sample analyzed in the laboratory (often less than 1 gram) to volumes of matrix that may be a million or more times larger. This concept is shown in Exhibit 2, which illustrates a sampling design where too few high-quality analytical samples were collected,

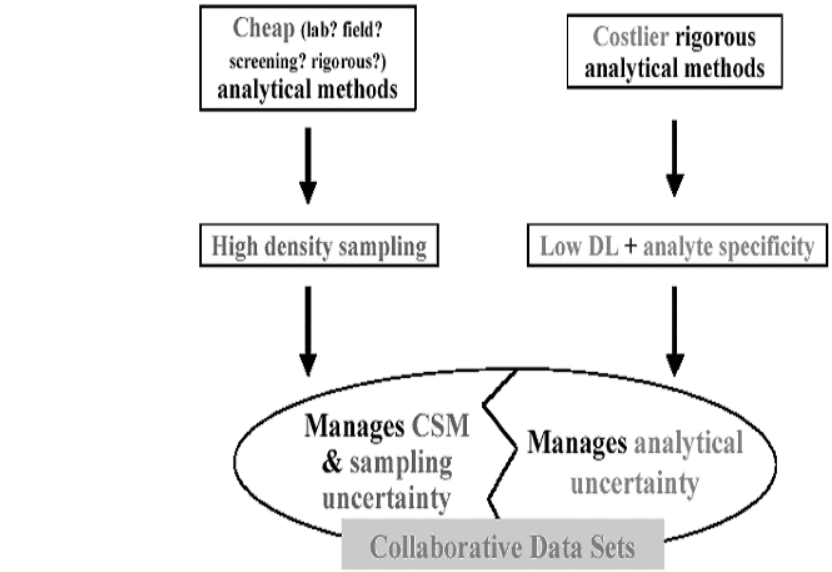

Exhibit 3. Collaborative data sets increase data quality in heterogeneous matrices (Crumbling et al., 2003)

causing important areas of contamination to be missed and the true extent of the contamination to be undefined. Conversely, the lower portion of the exhibit shows the same situation where many samples were collected, identifying and delineating the impacted areas. When the sampling density (number of samples per unit volume of environmental media) is insufficient to capture the effects of heterogeneity, incomplete or inaccurate CSMs are produced. Estimates of the nature and extent of contamination may be seriously biased, resulting in insufficient remedial designs that lead to more sampling and redesign.

The Triad approach recognizes that to develop an accurate CSM, high-density sampling is needed to understand the effects of site and matrix heterogeneity on contaminant distribution. However, not all the samples necessary to achieve the target sampling density need to be of the highest analytical quality. In fact, in most cases, high-quality analytical tests are seldom needed to understand contaminant distributions well enough to refine the CSM in the field; however, they are quite useful if it is advantageous to develop associative relationships (e.g., regressions) between field-derived and laboratory data sets so that all data types can be used in an integrated manner.

What is needed is a system that blends high-quality analytical methods with less expensive field methods in such a way that maximizes their respective strengths but compensates for their respective weaknesses for both real-time field conditions and postfield decision making and planning. A potential solution is to use field and fixed laboratory analysis in a collaborative data management program, as illustrated in Exhibit 3. Less expensive methods are used to increase sample density and build the CSM, and where unresolved analytical uncertainty remains, higher-quality analyses are added on samples for which their representativeness has already been determined. The purpose of this article is to describe an integrative conceptual framework for implementing collaborative data management programs in the field.

DATA USE CATEGORIES AND COLLABORATIVE DATA SETS

It would be ideal if all data could be "gold-plated"—that is, of sufficient rigor so that all site-related data would be suitable to address every conceivable decision supporting all potential land reuse options, from an unrestricted land use such as residential to a restricted land use such as industrial. However, as explained above in the case of cost and logistics, this ideal has proven to be impractical and exacerbated when time is factored in—two-week analytical turnaround times are not unusual for fixed laboratories. Despite poor control over sampling-related variables, the environmental community has labored under the paradigm that treats all data produced in a certified, fixed laboratory as of the highest quality and free from uncertainty, and therefore suitable for decision making for the most sensitive unrestricted land use. Although few question the quality of the data, high analytical quality is insufficient to overcome the sampling uncertainty created by data paucity. A corollary of this dysfunctional paradigm is the mistaken impression that data produced outside of certified, fixed laboratories are suspect, or worse, unusable.

This first-generation data quality model, focused as it is on analytical quality, discourages the use of real-time tools. Restricting these tools to "screening" uses only has significantly lengthened the learning curve for the environmental community to gain the experience needed to understand how to use these tools efficiently and defensibly. This has led to a pattern of costly, inefficient, and flawed site characterizations. Furthermore, the assumption that choosing a fixed lab analytical method will automatically produce data good enough for every possible use has encouraged the environmental community to avoid learning the skills needed to assess data uncertainty and its impact on decision making. The default assumption has long been that the quality of the data is governed simply by what analytical method is used, rather than an appreciation for the multitude of factors that actually impact data's ability to support specific project decisions. Therefore, the community's experience matching data quality to data use has been limited.

Although few question the quality of the data, high analytical quality is insufficient to overcome the sampling uncertainty created by data paucity.

In contrast to the prevailing paradigm, the second-generation data quality model used by the Triad approach acknowledges that *both* sampling and analytical variables impact data quality, and that data quality is best defined operationally according to specific data use. Our experience with Triad approach projects is that transitioning from the simple, but dysfunctional, first-generation data quality model to the more complex, but realistic, second-generation data quality model is difficult for many regulators and practitioners. It has been especially difficult to (1) integrate the idea of sampling uncertainty into a data quality classification scheme and (2) develop a scheme that treats nontraditional analyses (often field techniques) fairly. To facilitate Triad projects, a new framework for integrating data sets based on data use is necessary to allow the practical implementation of collaborative data sets.

In order to build a more confident CSM while maintaining analytical quality, Triad projects rely on a collaborative analytical scheme to increase sample throughput to support real-time decision making and to increase sampling density to control for sampling variability. To create a framework that matches the integration of sampling and analytical quality to a decision, it is useful to create data use categories that explicitly express the intended data use. The data use categories cannot be defined simply by the analytical technique used to generate the data. As noted above, the ana-

lytical technique alone is a poor predictor of the data use category, because the same technique may produce data that fall into any one of the different categories, or into all categories, even on the same project. For example, X-ray fluorescence (XRF) analysis for metals can produce data effective for decisions supporting the full range of restricted to unrestricted land use options. The appropriate use of XRF data depends on the analyte, method modifications, sample processing, matrix interferences, and, of course, the nature of the decision (contamination delineation, regulatory compliance, and risk assessment) to which the data are applied. Trying to define data quality according to the analytical technique or method and laboratory certification ignores too many important variables that could affect decision making if they were made explicit.

A data integration framework must take into account a number of factors that can limit how the data can be used to support science-based cleanup decisions.

A data integration framework must take into account a number of factors that can limit how the data can be used to support science-based cleanup decisions. These factors arise from the intersection of the analytical technique, site-specific matrix characteristics, the regulatory oversight structure, and project goals. These factors include:

- sample detection/quantitation limits and their relationship to the project's decision/action levels or thresholds;
- compound or analyte specificity of the method or instrument;
- sample support considerations for sample collection, preparation, subsampling, and extraction/digestion methods, as well as the determinative analytical method;
- regulator acceptance/certification status;
- the presence of matrix-specific physical or chemical interferences that impact sample collection, processing, and/or analysis; and
- the level of quality assurance/quality control (QA/QC) used to evaluate the performance of each component in the chain of sampling and analytical techniques.

Projects that need to integrate data of varying analytical rigor may need data integration frameworks that account for how they will be used. The example framework described below was developed to support decision making regarding contaminated soil in Brownfields projects. More testing by Triad practitioners is needed to know whether this scheme is general enough to work across widely different project scenarios. This article does not address data sets used to support design for engineered remedial systems. This article attempts to articulate how data use categories can support a collaborative sampling and analytical scheme to manage sources of uncertainty in data sets used for field-, compliance-, and risk-based decisions.

DATA USE CATEGORIES DESCRIPTION

Four data use categories, briefly described in Exhibit 4 and with more detail below, fall into two primary areas:

1) Data for rapid CSM development during dynamic field investigation programs that can be used to manage sampling uncertainty, as indicated by the left side of Exhibit 3:
   a. CSM:dirty
   b. CSM:clean

Data Use Categories That Support Rapid Development of CSM

Data Use Category	Application	Activity Supported	Limitations of Category Use	Conditions for Use
Build CSM: dirty data use category	CSM development for situations with elevated concentrations above action levels	Used to rapidly process high numbers of samples to advance csm and delineate impacts	Compound specificity or detection limits are insufficient to support regulatory decisions about "clean" areas	Normally applied with dynamic work strategy since primarily use real-time measurement devices
Build CSM: clean data use category	CSM development for situations with low concentrations at or below action levels	Define "clean" boundaries of impacted areas with confidence to identify areas/ volumes of the CSM	Compound specificity and detection limits are sufficient to meet action levels, but not sufficient to comply with regulatory certification	Applied in conjunction with CSM: dirty using a collaborative data management strategy and dynamic work strategy
Data Use Categories for Managing Analytical Uncertainty				
Polish CSM: compliance data use category	To satisfy regulatory requirements for full laboratory certification and strict adherence with method QA/QC and reporting deliverables	Effective for meeting site closure, no further action (NFA) and/or compliance monitoring expectations	No limitations on data-quality decisions with regard to analytical method; however, mature CSM is needed	Collaborative data management strategy: samples collected after CSM has evolved sufficiently to guide selection of appropriate locations
Polish CSM: risk-calc data use category	Most stringent from both scientific and regulatory perspective for quantitative risk assessment	Quantitative risk calculations to assess human health and ecological exposure from site chemicals	No limitations because of strict adherence to laboratory certifications and method QA/QC	Collaborative data management strategy: sample locations identified after CSM has matured sufficiently to define exposure pathways and populations

Exhibit 4. Summary of data integration and categorization framework

2) Data for final refinement ("polishing") of the CSM by managing analytical uncertainty relevant to regulatory and risk-based decisions that require analyte-specific, analytically unbiased data sets, as indicated on the right side of Exhibit 3:
   a. CSM:compliance
   b. CSM:risk-calc

In actuality, contaminant data are not the only kind of data used to develop CSMs. All forms of site information, from chemical data to the site history to geotechnical and geophysical data, should be used to construct and refine the CSM. There can also be considerable overlap in how data sets are used. Cost-effectiveness and project efficiency are increased as data sets are designed to serve more than one use. However, for the sake of the proposed categorization described below, only the primary role of the data will be highlighted.

These four categories form the basis for an integrative framework that can be used to facilitate pre-field-planning discussions and management of field and fixed laboratory data produced in real time. They can also be used to decide how data will be integrated.

Cost-effectiveness and project efficiency are increased as data sets are designed to serve more than one use.

DATA USE CATEGORY—EFFECTIVE FOR CSM DEVELOPMENT FOR SITUATIONS WITH ELEVATED CONCENTRATIONS ABOVE ACTION LEVELS (CSM:DIRTY)

General Description

This data use category includes data sets that are suitable for detecting, delineating, and modeling higher concentration ("dirty") areas (those believed to be contaminated or suspected to exceed the established decision threshold) in real time. Using high sample throughput rates, this data use category is intended to support high-density sampling to rapidly advance the development of the CSM. Typically, only higher analyte concentrations are reported with confidence, generally limiting the use of this category of data when decisions require lower concentration limits. This is because certain important aspects of the analytical rigor and method performance (commonly, the quantitation limit and/or analyte specificity) may be insufficient to delineate lower concentration ("cleaner") areas with sufficient confidence to support regulatory decisions. Despite their lower rigor, these data are of known quality (i.e., the associated QC demonstrates that sampling and analysis are under control and adequate to support the intended data use). Whether they were generated by a certified laboratory/operator (where such certification programs exist) or not is not a determining factor. Other possible data uses within this category are instances where one analyte is used as a surrogate to indicate the presence and approximate concentration of other analyte(s) because there is an associative (although not necessarily a statistical regression) relationship between them. The CSM:dirty category applies when the predictive relationship is confident enough to predict areas/volumes that likely exceed the established action threshold, but is not confident enough to predict nonexceedances.

These data are generally produced by high throughput sampling and analysis procedures that allow large numbers of samples to be processed and reported in real time. The sample support varies from larger (if some form of sample compositing or large volume mixing is used) to quite small (e.g., in situ sensor systems). Small sample supports can be very useful for delineating discrete contaminant populations and distribution intervals, but this must be balanced against the potential for "nugget" effects (isolated small pockets of contamination) that increase the variability in a data set and can bias data in a nonrepresentative way. Increasing the number of readings to understand whether microheterogeneity is a problem for the matrix and analyte under investigation may control the uncertainty introduced by small sample supports.

Benefit of CSM:dirty Data Use Category

The chief benefit of the CSM:dirty data use category is to allow rapid development and refinement of the CSM for areas of higher concentration. Although the data are not generally effective for delineating "clean" zones, the information provided allows estimation of the number of distinct populations and a coarse estimate of variability in those populations across the site. Another way to state this is to say that the CSM:dirty category helps the project team to rapidly evolve the CSM through an understanding of contaminant distributions. This understanding is necessary to support project decisions that require the following inputs:

- Identify significant sampling variables and the mechanisms to control for those variables as needed to manage intolerable decision uncertainty.
- Detect the presence of spatial patterning and determine whether that patterning is associated with known or suspected contaminant release, migration, or partitioning mechanisms.
- Accurately estimate volumes of contaminated material to evaluate treatment and disposal options and predict remedial costs.
- Identify and evaluate exposure pathways.

This data use category is normally applied in conjunction with a dynamic work strategy since the sampling and analytical methods and instruments used are, for the most part, real-time measurement devices.

Determination of When Data Fall into the CSM:dirty Data Use Category

There are a variety of mechanisms by which data fall into the CSM:dirty data use category. Generation of data suitable only for modeling higher contaminant concentrations can be deliberate or inadvertent.

Generation of data suitable only for modeling higher contaminant concentrations can be deliberate or inadvertent.

- **Deliberate generation** of data occurs when sample processing and analytical techniques are selected for the express purpose of rapidly processing high numbers of samples. Although a data set that can be used for decisions at the lowest possible action level is always desirable when feasible, the limitations of technology, time, budget, personnel, bench space, or other constraints may make investment in more expensive or labor-intensive sampling and analytical options prohibitive. Most importantly, during systematic planning, it is determined that the type of uncertainty these data will manage in the given decision scenario can be adequately addressed with a less rigorous data set, so more expensive options are not necessary.
- **Inadvertent generation** occurs when a more rigorous data set (e.g., low detection) was planned and expected, but data quality ended up not being as "good" as the project team had hoped when the data came back from the lab. Although analytical quality is not what was planned, and it is now found inadequate for more stringent data uses (such as demonstrating regulatory compliance or calculating risk), the data can still have limited utility for building confidence in the

CSM. In other words, data that must be rejected for one data use may still be quite useful for another data use, so it may not need to be totally discarded. Remember, under the Triad approach, building the CSM is a critical activity, and many data have some use for that purpose as long they are of known quality. Reasons why CSM:dirty data may be inadvertently generated include:

1. Matrix interferences: Sample effects degrade the performance of methods that were expected to produce more rigorous data sets. An example is when a laboratory dilutes a sample extract to reduce interferences but inadvertently raises all or some of the target analytes' quantitation limits above their respective action levels.
2. Errors in planning: A standardized analytical method is used. Project planners assumed that the data quality would automatically be adequate for all possible decisions as determined after the data were collected. But no one noticed that the method was not really appropriate for all of the target analytes. For example, the planning team may not notice that the standard method has a quantitation limit set too high to establish "clean" for an analyte important to the project. Not until the results come back from the laboratory does the data user realize that the standard method was not designed to meet project needs. Proper planning would have determined that an alternate or modified method was needed.
3. Operator error: The operator/analyst may err by not following the project's standard operating procedures (SOPs) for sampling and analysis. Alternatively, the operator may be following the SOPs but fail to notice or report to management that the SOPs were poorly matched to the actual needs of project implementation. Problems with SOPs should be brought to the project manager's attention so that corrective action can be taken to avoid wasting resources on inappropriate data collection.
4. Instrumentation problems: Quality control may indicate that instrument, blank, or batch problems exist that unexpectedly limit the utility of data for decision-making purposes.

The operator/analyst may err by not following the project's standard operating procedures (SOPs) for sampling and analysis.

DATA USE CATEGORY—EFFECTIVE FOR CSM DEVELOPMENT FOR SITUATIONS WITH LOW CONCENTRATIONS AT OR BELOW ACTION LEVELS (CSM:CLEAN)

General Description

This data use category includes data of sufficient analytical quality to delineate areas where contaminant concentrations are lower than regulatory limits (i.e., "clean" or "compliant" soil). The purpose of this data use category is to identify areas/volumes of the CSM that depict "clean" so that these areas can be bounded with confidence. Commonly, the determining factor for the data use is whether quantitation limits are low enough. But other aspects of data quality, such as freedom from interferences, bias, and precision, may also be determining factors. Generally, a data set that is sufficient to model populations of clean matrix will also be reliable for modeling populations of more contaminated matrix, but there can be exceptions to this. For example, a highly sensitive technique that works well on simpler, low-concentration matrices may be sub-

ject to interference and produce false positive or false negative detections when challenged with a real-world complex matrix with high concentrations of pollutants. Site history and knowledge of likely interferences can be an important factor when assigning a data use category.

The intention of the CSM:clean data use category is, like the CSM:dirty category, to use real-time measurement systems to create the information that will drive a dynamic work strategy to cost-effectively evolve the CSM. Ideally, methods and instruments used in the previous category (CSM:dirty) work in concert with methods and instruments in this category (CSM:clean) to build a collaborative data set. Thus, it is critical during the systematic planning process to develop the respective decision logic and rules that will guide the field team to use data generated by these two categories.

Benefits of the CSM:clean Data Use Category

Although the data sets within the CSM:clean data use category are effective for identifying the location and boundaries of clean areas of the CSM, the data quality, from either a sampling or analytical standpoint, may not be sufficient for more rigorous data use (such as regulatory compliance leading to a decision of No Further Action). This may stem from the use of techniques that are nonspecific or are modifications of standard methods that have not been widely accepted. For example, a technique may report contaminant groups or classes but cannot supply the analyte-specific concentrations needed for many quantitative data uses. Even if the results are analyte-specific, the degree of bias or imprecision in the data set may be known to exceed that needed for more stringent data uses. Sometimes, slight modifications are made to a standard method in order to increase sample throughput. Under current laboratory certification procedures, any method modifications may be unacceptable to the certifying authority, even if the analytical performance is unchanged or improved. Even when the scientific decision-making value of the data remains unchanged, regulatory rejection of the data for risk or compliance uses may force the data be restricted to CSM:clean and CSM:dirty uses only. As the cleanup industry evolves to focus on decision uncertainty management, we hope that such regulatory restrictions will be reconsidered in the interests of promoting efficient, effective investigations and cleanups.

Under current laboratory certification procedures, any method modifications may be unacceptable to the certifying authority, even if the analytical performance is unchanged or improved.

Despite these limitations, CSM:clean data sets are of great value to the project by:

- reducing the cost of generating high sampling densities;
- creating the real-time availability of results to support a dynamic work strategy; and
- serving as a check of the performance and back-up of findings for CSM:dirty data use category methods and instruments.

These data are effectively used to stratify populations for statistical purposes or locate and delineate areas/volumes requiring no further action. As with all data used in a Triad project, they are data of known quality (i.e., the in-field QC establishes their adequacy to support data use). Whether or not a certified service provider (where such certification programs exist) generates the data is not a determining factor. Along with the CSM:dirty category, these data are generally used as collaborative data that manage sampling uncertainties (Exhibit 3).

Determination of How to Define Data Used in This Category

Like the CSM:dirty category, CSM:clean data sets may be generated deliberately as part of the project plan or inadvertently due to complications from matrix interferences, sampling uncertainties, human error, or instrument QC problems that compromise the reliability of data that were intended to be more rigorous applications. Data can be expected to fall into the CSM:clean data use category under the following conditions:

. . .quantitative results may have only limited utility because they are known to be significantly biased or imprecise due to sampling or analytical limitations.

- The analytical technique reports only compound class-specific (not analyte-specific) data. Or, if the technique reports analyte-specific results, results are reported qualitatively (i.e., greater or less than a certain value) or semiquantitatively (i.e., as concentration ranges). Alternatively, quantitative results may have only limited utility because they are known to be significantly biased or imprecise due to sampling or analytical limitations. Despite the data uncertainty, the data are entirely suitable for supporting selected decision scenarios because they are of known quality and detection limits are below appropriate action levels. For example, non-detect or low-detect data may be highly predictive for an entire class of compounds to which the technique responds and may lead to a high level of confidence to render a decision on cleanliness.
- Data for one analyte can be used as a surrogate to indicate the presence and approximate concentration of another analyte(s) because there is a sufficiently strong predictable relationship between them at lower concentrations to confidently predict when matrix concentrations are not exceeding applicable action levels.
- Data may also be relegated to this category if regulatory programs have certification/accreditation or other requirements that limit regulatory acceptance of data generated using nontraditional methods, even if the data would be considered acceptable for the intended use from a purely scientific standpoint.

DATA USE CATEGORY—EFFECTIVE FOR MANAGING ANALYTICAL UNCERTAINTY FOR THE PURPOSES OF REGULATORY COMPLIANCE (CSM:COMPLIANCE)

The compliance data use category includes data sets that are effective for meeting regulatory site closure or compliance monitoring expectations for reporting limits, analyte specificity, precision, bias, and certification/accreditation of the service provider. These data sets "polish" the CSM by managing any lingering analytical uncertainty with respect to contaminant identity and low-level concentrations. Normally, these data are produced in strict adherence to a particular regulatory agency or group's data quality and reporting requirements (data deliverables).

Since the majority of analytical techniques used to generate these data use very small sample supports (e.g., 1 to 10 grams of soil) in comparison to the mass of the parent matrix to which the results will be extrapolated, uncertainty about the representativeness of the analytical sample may be very high. Often, the dynamic work strategy for this data category will call for a percentage of split samples that are to be analyzed at fixed laboratories and in the field. The results of the split samples help establish the confidence in the fuller data set as well as help to refine the CSM, which would have been developed from the CSM:dirty and CSM:clean data use categories. The percentage of

split samples should be guided by several considerations. If data use involves decision making at an action level, a large number of split samples should be focused on managing the decision uncertainty around that action level. If the planned data use warrants it, split samples may also be used to develop the appropriate statistical regressions between field and fixed data, in which case split samples need to be taken across the concentration range relevant to the decision. Either case requires knowing the approximate concentration range of the sample before selecting it for split-sample analysis. Random selection of an arbitrary percentage of samples is likely to produce a data set with a high number of non-detects, which would not be useful for data comparison and statistical purposes (Interstate Technology and Regulatory Council, 2003). Samples for the CSM:compliance data use category should be collected after the CSM has evolved sufficiently to guide the selection of appropriate locations and number of samples. With some advance planning on sample volume, samples can be field-tested and archived for later split-sample analysis as necessary. For example, nondestructive testing such as handheld XRF technology would allow for field characterization and future laboratory analysis. Compliance-use data are distinguished from risk-calculation data because the demands on quantitation limits and data precision and bias can be less stringent for determining compliance with a regulatory threshold than for what would be required for quantitative risk calculations.

DATA USE CATEGORY—EFFECTIVE FOR MANAGING ANALYTICAL UNCERTAINTY FOR THE PURPOSES OF QUANTITATIVE RISK CALCULATIONS (CSM:RISK-CALC)

The risk-calc data use category is the most stringent from both a scientific and regulatory standpoint. The policy implications of risk assessments generally demand that any applicable laboratory/operator certification requirements be met. Quantitative risk assessment requires low quantitation limits and analyte specificity. In addition, sitewide representativeness of the data is essential to develop a true picture of risk for the site—particularly in the case of probabilistic risk assessments. Biased data, such as those from AOC-driven sampling regimes, may preclude sitewide risk assessments. A simple example is a site where only a portion is contaminated, say 50 percent. If the data set were composed solely of samples taken from the AOC portion of the site, the risk assessment would likely overestimate the risk by a factor of two.

The policy implications of risk assessments generally demand that any applicable laboratory/operator certification requirements be met.

From a data standpoint (not just an analytical standpoint), low bias and good precision are required in the actual data set (which includes the impact of heterogeneity, not just bias and precision of the analytical technique) to reduce uncertainty in data that could be used for exposure estimates across the designated exposure unit. These demands on data rigor make it desirable to obtain the best analytical quality that is technically feasible. As with the CSM:compliance data use category, the analytical techniques used to generate risk-calc data typically use small sample supports; hence, sampling uncertainty will be high unless sampling variables are strongly managed at both macro (sampling locations) and micro (sample preparation) levels. A confident and mature CSM capturing any significant contaminant patterning and variability should be the basis for selecting sample number and locations and associated collection and handling procedures. Both between- (i.e., macro) and within-sample heterogeneity (i.e., micro) should be measured and controlled so these expensive (produced in strict compliance with

standard method QA/QC requirements, certifications, and data reporting deliverables) data points will have maximal effectiveness for the risk-calc data use.

The more stringent data use categories (compliance and risk-calc) are supported when the appropriate sampling and analytical techniques were selected and implemented and no confounding analytical interferences, operator error, cross-contamination, or QC problems occur that could cause the data to be flagged/qualified. These data sets have historically been generated through rigorous sampling and analytical procedures performed in the controlled environment of a field or fixed laboratory that can ensure proper equipment maintenance, calibration, sample processing, and storage. At this point in time, this generalization is still largely true, but exceptions are growing as laboratory instrumentation is miniaturized and made more rugged for field deployment. One such example is field-portable gas chromatography/mass spectrometry (GC/MS) instruments equipped with standardized sample preparation modules.

Extreme caution must always be exercised whenever very small subsample supports are used.

Typically, *ex situ* samples are required because of the need to control sample support and particle size when generating concentration data that can be appropriately compared to regulatory action levels or risk-derived decision thresholds. Extreme caution must always be exercised whenever very small subsample supports are used. Strict control over sample homogenization, preparation, and subsampling procedures are required in order to reduce subsample variability, produce data that is comparable across different analytical techniques, and ensure that the correct matrix population has been targeted for representative analysis (see EPA/600/R-03/027, Nov 03, http://www.epa.gov/esd/tsc/images/particulate.pdf).

Since it is generally more expensive to generate data suitable for the more stringent data uses, samples should be carefully selected to achieve the highest degree of information return for the money being spent. Samples should be of known representativeness (i.e., the mature CSM and the decision framework should establish the appropriate sample location and population to be targeted for analysis). Because these data sets are costly, they are reserved for managing analytical uncertainty that cannot be managed in a less expensive way. Control over sampling uncertainty and development of the CSM is performed (when at all possible) using less expensive options that support high data density and delineation of the populations to be targeted for risk or remedial decisions. Once those populations are defined, samples representative of those target populations may be collected for more expensive procedures used to create an unrestricted and unlimited land use–driven decision-making data set. This concept is illustrated in Exhibit 5, where the bulk of site sampling used to rapidly grow and mature the CSM is within the CSM:dirty and CSM:clean data use categories. As the confidence in the CSM improves, the ability to select those samples that require more stringent analytical rigor is driven by the objectives of the project and the mature CSM. Both CSM:compliance and CSM:risk-calc data sets should be designed to manage whatever relevant analytical uncertainties remain after target contaminant populations have been defined, as indicated in Exhibit 3.

SUMMARY

Judicious blending of different data use categories maximizes the return on investment in a site characterization project because sampling uncertainty, and other important sources of erroneous information, are identified and controlled. A carefully designed

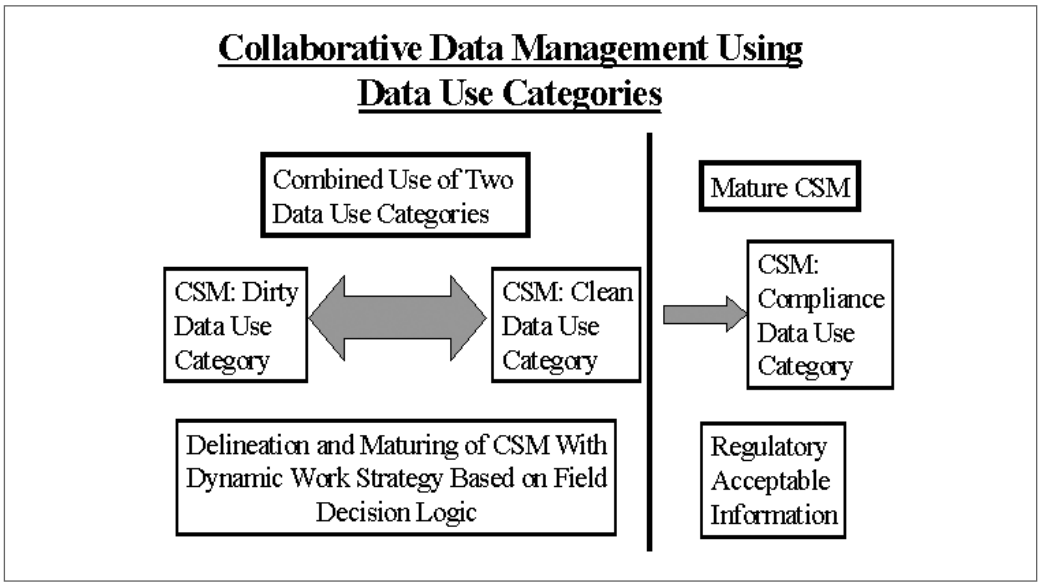

Exhibit 5. Collaborative data management using data use categories

combination of sampling and analytical techniques manages the fundamental mismatch between the tiny volume of traditional analytical samples and the very large volume of material to which analytical results are extrapolated. When dynamic work strategies are used in conjunction with collaborative data management systems, sample patterns can be more extensive without ignoring concerns over known contaminated areas while overcoming the tendency to focus on AOCs or other biasing mechanisms.

High-density sampling targeted to areas of decision uncertainty is essential to account for site heterogeneity and for understanding contaminant distributions at scales ranging from macro (between samples) to micro (within a single sample) for both time and space. Without this understanding, for example, the representativeness of isolated 1-gram analytical samples is unknown. Is there any confidence that the analytical result from a 1-gram sample truly represents the average concentration in the sample jar? Is it legitimate to extrapolate that concentration result to represent the average concentration for the 500 cubic yards of soil in the field grid from which that tiny sample was taken? When the representativeness of data is unknown, the data quality is unknown, no matter how much analytical quality control was performed.

Cost-effective and efficient site remediation and scientifically defensible decisions require accurate site characterization. The CSM is at the center of site characterization. It is the hub that guides iterative rounds of sampling to detect and bound spatial patterns that indicate sources, exposure pathways, and hot spots. The iterations required to refine a CSM are most cost-effective when performed in real time. Therefore, data used to test and refine the CSM must be available in real time. A collaborative data management system, when used by a skilled and experienced field team, is the only procedure currently available for collecting enough site information to mature the CSM in a timely manner. An easily understandable framework to structure data quality in a way that supports scientifically defensible decisions and efficient projects has been proposed. This framework can also be used to meet regulatory oversight objectives.

REFERENCES

American Society for Testing and Materials (ASTM). (1998a). Standard guide for accelerated site characterization for confirmed or suspected petroleum releases. ASTM E-1912-98. West Conshohocken, PA: Author.

American Society for Testing and Materials (ASTM). (1998b). Standard practice for expedited site characterization of vadose zone and ground water contamination at hazardous waste contaminated sites. ASTM D-6235-98. West Conshohocken, PA: Author.

American Society for Testing and Materials (ASTM). (1999). Guide for developing and implementing short term measures or early actions for site remediation. ASTM D-5745-95. West Conshohocken, PA: Author.

Applegate, J. L., & Fitton, D. M. (1997). Rapid site assessment applied to the Florida Department of Environmental Protection's Drycleaning Solver Cleanup Program. In Proceedings of the Superfund XVIII Conference (Vol. 2, pp. 695–703), Washington, DC.

Burton, J. C., Walker, J. L., Aggarwal, P. K., & Meyer, W. T. (1995). Expedited site characterization: An integrated approach for cost- and time-effective remedial investigation. Argonne, IL: Argonne National Laboratory.

Connecticut Department of Environmental Protection, Leaking Underground Storage Tank Program. (2000). Expedited site assessment: The CD (Version 1.0)—UST site investigation guidance for a new millennium. Hartford, CT: Author.

Crumbling, D., Groenjes, C., Lesnik, B., Lynch, K., Shockley, J., VanEe, J., et al. (2001). Applying the concept of effective data to contaminated sites could reduce costs and improve cleanups. Environmental Science and Technology, 35(19), 405A–409A.

Crumbling, D. M., Griffith, J., & Powell, D. M. (2003). Improving decision quality: Making the case for adopting next-generation site characterization practices. Remediation, 13(2), 91–111.

Ellerbusch, F., Mack, J., & Shim, J. S. (2004). Using the Triad approach to expedite the acquisition of an Abbott District School Site. Remediation, 14(2), 85–105.

Interstate Technology and Regulatory Council (ITRC). (2003). Technical and regulatory guidance for the Triad approach: A new paradigm for environmental project management (SCM-1). Prepared by the ITRC Sampling, Characterization and Monitoring Team. Washington, DC. Retrieved October 20, 2004, from http://www.itrcweb.org/SCM-1.pdf.

Mack, J., Ellerbusch, F., & Librizzi, W. (2003). Characterizing a Brownfields recreational reuse scenario using the Triad approach—Assunpink Creek Greenways Project. Remediation, 13(4), 41–59.

Robbat, A. (1994, May). Case Study Fort Devens, Massachusetts: Methods development, feasibility, cost/benefit analysis for performing on-site thermal desorption gas chromatography/mass spectrometry of organic compounds at Army facilities. Paper prepared for the U.S. Army Environmental Engineering Center, Aberdeen, MD.

Robbat, A. (1997). Dynamic workplans and field analytics: The keys to cost effective site characterization and cleanup. Medford, MA: Tufts University Center for Field Analytical Studies and Technology.

Texas Natural Resource Conservation Commission. (1995). Accelerated site assessment process procedure: A guidance manual for accessing LPST sites in Texas. Austin, TX: Author.

United States Environmental Protection Agency (US EPA). (1997). Expedited site assessment tools for underground storage tank sites: A guide for regulators. EPA 510-B-97-001. Washington, DC: Office of Underground Storage Tanks, Office of Solid Waste and Emergency Response.

United States Environmental Protection Agency (US EPA). (2000). Innovations in site characterization. Case study: Site cleanup of the Wenatchee Tree Fruit Test plot site using a dynamic work plan. EPA-542-R-00-009. Washington, DC: Author.

United States Environmental Protection Agency (US EPA). (2001). Improving sampling, analysis, and data management for site investigations and cleanups. EPA-542-F-01-030a. Washington, DC: Office of Solid Waste and Emergency Response, U.S. Environmental Protection Agency.

Woll, B., Mack, J., Ellerbusch, F., & Vetter. J. (2003). Facilitating Brownfields transactions using Triad and environmental insurance. Remediation, 13(2), 113–130.

Jim Mack is a director at the Northeast Hazardous Substance Research Center located in the York Center for Environmental Engineering and Science at the New Jersey Institute of Technology (NJIT), Newark, New Jersey. He holds a BS in geology from Waynesburg College and an MS in earth science from Adelphi University.

Deana M. Crumbling has worked in the hazardous waste site cleanup arena over the past 12 years, and has been at the US EPA since 1997. She is an analytical chemist with clinical, industrial, and research experience. She holds a BS in biochemistry, a BA in psychology, and an MS in environmental science.

Fred Ellerbusch, P.E., DEE, is a director at the Northeast Hazardous Substance Research Center located in the York Center for Environmental Engineering and Science at NJIT, and a faculty member at the University of Medicine and Dentistry of New Jersey (UMDNJ) School of Public Health. He is a PhD candidate and holds an MPH from UMDNJ as well as an MS in environmental engineering and a BS in civil engineering from NJIT.
