This is a preprint of an article published in *Remediation: The Journal of Environmental Cleanup Costs, Technologies & Techniques, Vol. 15*, No. 1, pp. 21-37. The published reprint is available through http://www.interscience.Wiley.com/

A Data Integration Framework to Support Triad Projects

James Mack, New Jersey Institute of Technology - Northeast Hazardous Substances Research Center

Deana Crumbling, USEPA Office of Superfund Remediation and Technology Innovation

Fred Ellerbusch, New Jersey Institute of Technology - Northeast Hazardous Substances Research Center

ABSTRACT

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 to support scientifically defensible decisions and efficient projects. The currently used data quality framework focuses heavily on analytical quality, but is relatively silent on the subject of how sampling uncertainties impact data quality. Since it is well known that contaminant heterogeneity introduces several sampling-related variables that critically impact data quality, a successful data quality framework must be able to account for the contributions of both sampling and analytical uncertainty to data quality. Because of the wide variety of decisions faced by site cleanup professionals, a constructive data quality framework must also anchor data quality assessment in the specific needs of the decision-making process. This paper explores such a framework.

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 lifecycle of characterization, remediation, and site reuse (Exhibit 1; Crumbling, et al 2001; USEPA, 2004). Systematic planning is the period in the project when the initial Conceptual Site Model (CSM) is developed. A Triad-based CSM serves as a mechanism to achieve consensus understanding among the various stakeholders about the nature and extent of contamination, potential receptors, potential risk mitigation measures, and ultimately, the most satisfactory disposition of the site. Systematic planning is the time when project end points are clearly articulated, the range of remedial approaches defined, clean up criteria established, and the general investigation methods and procedures are developed.

This is a preprint of an article published in *Remediation: The Journal of Environmental Cleanup Costs, Technologies & Techniques, Vol. 15*, No. 1, pp. 21-37. The published reprint is available through http://www.interscience.Wiley.com/

Dynamic work strategies are integrated systems of decision logic and decision 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. Decision logic and rules are developed by consensus among the project team and stakeholders during project planning. 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 geophysical characterization, and location information (geographic position system), etc., as well as data management and display software. Any data generated in the field are useable as long they are of known quality (i.e., quality control (QC) is within performance limits), are appropriate to the decision at hand, and are based on prior stakeholder agreement.

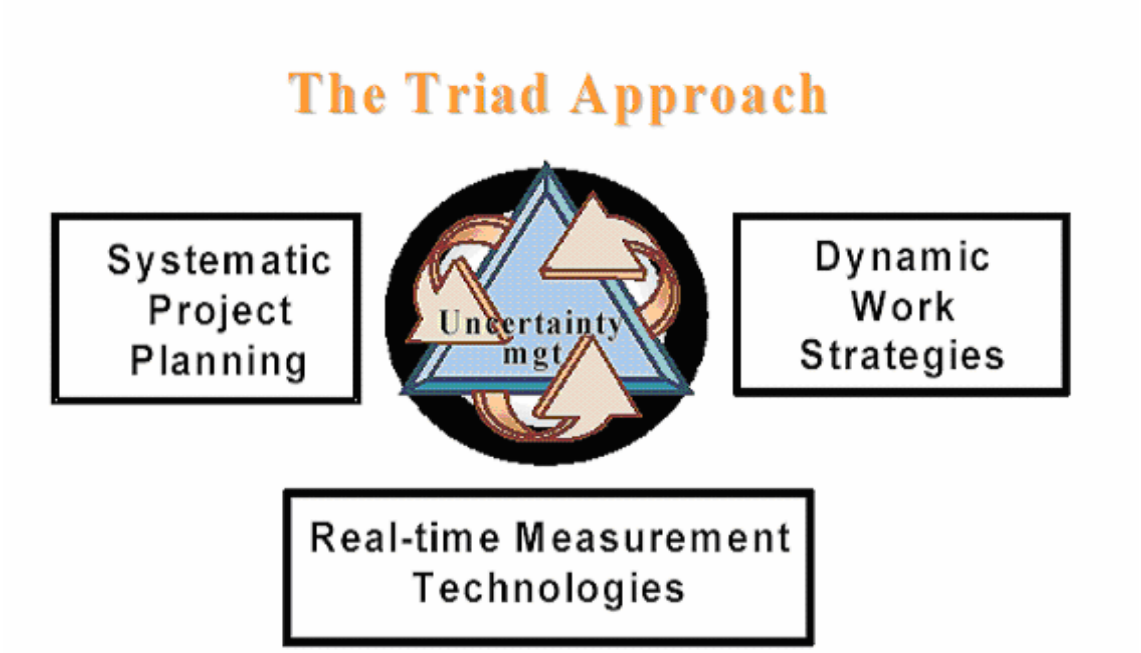

Exhibit 1. The Triad approach components (USEPA, 2004)

Literature on the use of innovative site management strategies, including the Triad Approach, is emerging in the form of research, guidance from government and standards-setting organizations, and case studies (For more information about innovative site management strategies see Applegate & Fitton, 1997; ASTM, 1998; ASTM, 1998a; ASTM, 1999; Burton et al, 1995; Crumbling et al, 2001; CT-LUSTP, 2000; Ellerbusch et al, 2004; Mack et al, 2003; Robbat, 1994, 1997; TNRCC, 1995; USEPA, 1997, 2000, 2001; Woll et al, 2003).

This is a preprint of an article published in *Remediation: The Journal of Environmental Cleanup Costs, Technologies & Techniques, Vol. 15*, No. 1, pp. 21-37. The published reprint is available through http://www.interscience.Wiley.com/

The large volume of data produced through a Triad Approach investigation generally demands that a pre-defined data management system be used to store and process data results in real-time. The data management system should be constructed to accommodate the overlapping collaborative methods and procedures that will be used to manage both sampling and analytical aspects of data uncertainty. The data management system is also useful for managing the QC data generated during implementation of the project's quality assurance project plan (QAPP).

In addition to the QC exploited by traditional projects, the real-time nature of Triad projects offers three additional quality control mechanisms that greatly increase project efficiency:

1) Real-time review of QC results to evaluate whether the sampling and analytical techniques are performing as required (i.e., are in control), with immediate intervention if they are not;
2) Real-time adjustment or focusing of QC procedures and frequencies to quickly identify and reduce data uncertainty in relation to project decisions as the CSM evolves and site conditions are better understood; and
3) Real-time comparisons between newly acquired real-time data and the current CSM to detect incompatibility or conflict.

Real-time detection of analytical performance issues minimizes the time and resources spent generating and managing useless data. Real-time focusing of QC means that QC is fine-tuned to address decision uncertainty as the project unfolds, so that precious dollars are spent only on those QC parameters that most benefit the project. Iterative real-time evaluation of compatibility between the CSM and new data as it is being generated serves as a powerful quality assurance mechanism. Conflict between the data and the CSM triggers an appraisal 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).

MATRIX HETEROGENEITY AND DEVELOPING ACCURATE CSMs

Environmental matrices are heterogeneous at both the macro and micro scales. Generating data sets that accurately represent important spatial heterogeneities is critical to developing a CSM that reflects relevant site conditions. Although sample data may be correct in the sense that the analytical results are accurate for the tiny amount of material analyzed, extrapolating those results to the much larger volumes encompassed by sub-units of a site, or the entire site, can easily 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. Misleading CSMs can lead directly to erroneous assessment of risk or faulty decisions about the remedial strategy.

Under the traditional approach to site characterization, over-reliance on an Area of Concern (AOC) approach and high quality (but expensive) analyses permits too few samples to be collected, risking generation of a faulty CSM. There are a multitude of underlying reasons and reinforcing factors that can lead to this problem, including:

- AOCs are by definition biased toward contamination, so the volume of material that is actually compliant may be underestimated, causing cleanup and redevelopment cost estimates to be needlessly inflated.

This is a preprint of an article published in *Remediation: The Journal of Environmental Cleanup Costs, Technologies & Techniques, Vol. 15*, No. 1, pp. 21-37. The published reprint is available through http://www.interscience.Wiley.com/

- Representative site-wide sampling plans have sometimes conflicted with AOC-focused sampling plans.
- More expensive high quality analytical techniques reduce the number of samples in a given sampling budget.
- Regulation-based characterization regimes, sometimes emanating from an AOC-driven approach, are often implemented as prescriptive sampling and analysis strategies that leave little room to adjust to project-specific needs and site-specific conditions.

When too few samples are collected, there is little choice but to extrapolate the results of the tiny samples analyzed in the laboratory (sometimes less than one gram) to volumes of matrix that may be billions of times larger. High concentration results may be assumed to represent the concentration of large volumes or areas of matrix, implying that a large contaminant mass is present. But the actual volume of contaminated matrix could actually be quite small. Similarly, low concentration results may be inappropriately interpreted to assume that large areas are "clean," when that is not true. Both decision errors lead to inefficient projects that waste time and money. The later condition is illustrated in the upper panel of Exhibit 2, which depicts a sampling design where too few analytical samples were collected. Despite the high quality of the analyses, important areas of contamination are missed and the true extent of the contamination is inaccurately defined. Whenever the sampling density (number of samples per unit volume of environmental media) is insufficient to capture the effects of heterogeneity, there is a high likelihood that incomplete or inaccurate CSMs will be produced. Estimates of the nature and extent of contamination may be seriously biased, resulting in insufficient remedial designs that lead to more sampling and remediation as errors are discovered. In contrast, the lower panel of Exhibit 2 illustrates the more accurate CSM developed when sufficient samples can be collected to detect and accurately delineate impacted areas.

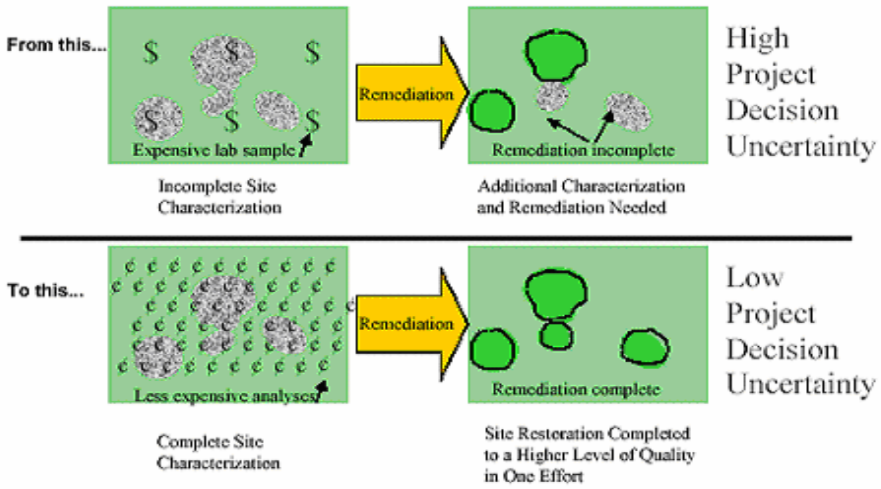

Exhibit 2. Sample Representativeness and Uncertainty. By collecting a larger number of less expensive (¢) samples a more complete understanding of site conditions can be achieved than by expensive high quality analytics ($) alone (ITRC, 2003).

This is a preprint of an article published in *Remediation: The Journal of Environmental Cleanup Costs, Technologies & Techniques, Vol. 15*, No. 1, pp. 21-37. The published reprint is available through http://www.interscience.Wiley.com/

The Triad approach recognizes that to develop an accurate CSM, high-density sampling is needed to understand contaminant distributions and the effects of heterogeneity on sample results. "Sampling density" may refer to the number of samples per area or volume of matrix, or to the number of samples used to support a particular decision. Samples per matrix volume can refer both to the number of discrete samples taken from different locations throughout the matrix, and to the number of repeat measurements taken from (for example) the same jar of soil to characterize within-sample variability. Triad also recognizes that different areas of the same site will probably need different sampling densities. Higher densities target those areas where decisions are the most uncertain. This goal can be achieved without needing to analyze all samples by the most expensive, high quality methods. In most cases, contaminant patterns can be understood and the CSM refined well enough to delineate "no-further-action" vs. "action-required" areas without having to analyze all of the samples by definitive analytical methods. However, judicious application of high quality analytics is generally required to supplement less expensive testing for the purpose of developing associative or predictive relationships that permit the less rigorous analyses to be interpreted correctly. Integrating a variety of analytical options and data types to build, refine, and polish the project's CSM is termed "collaborative data management" by Triad practitioners.

Since field work is most efficient when it can be adjusted in real-time, inexpensive methods that generate data *in the field* are extremely useful. But traditional notions of data quality and inappropriate usage tarnished the reputation of field methods through the 1990s, limiting their acceptance despite clear evidence that projects can be more efficient and less costly when high density field analyses were used. If we are to overcome the institutional hurdles now faced by field methods, a data quality framework is needed that ensures that field methods are used appropriately. That framework would explicitly blend high quality analytics with less expensive data generation techniques in ways that maximize their respective strengths but compensate for their respective weaknesses. These mutually supporting data sets would allow maximal utility of all kinds of data for purposes of efficient, real-time site characterization, as well as site reuse design, risk assessment, remediation, and compliance. The solution is to use field and fixed laboratory analyses in a collaborative data management strategy as illustrated in Exhibit 3 on the next page.

Collaborative data sets rest on the concept that less expensive methods should be used to increase sample density and build the CSM—a prime requisite to establish the representativeness of any individual sample (i.e., the ability to confidently extrapolate the results of 1- or 10-gram analytical samples to "represent" larger volumes of matrix). Where unresolved analytical uncertainty remains, higher quality analyses are then performed on samples for which the representativeness is already established. This paper posits a need for a conceptual framework to support collaborative data management programs both in the field and in regulatory programs.

DATA USE CATEGORIES AND COLLABORATIVE DATA SETS

It would be ideal if ALL data could be "gold-plated," that is, of maximum possible quality suitable for every conceivable decision under all potential land reuse options. However, as explained above, cost and logistics make this ideal impossible. Sole reliance on the highest analytical quality is also impractical when time is factored in – two-week analytical turn-around times and greater are not unusual for fixed laboratory analysis, hindering chances for real-time field decisions. Yet the environmental community has labored for two decades under a paradigm that treats data produced in a certified, fixed laboratory as if it is gold-plated, automatically of the

This is a preprint of an article published in *Remediation: The Journal of Environmental Cleanup Costs, Technologies & Techniques, Vol. 15*, No. 1, pp. 21-37. The published reprint is available through http://www.interscience.Wiley.com/

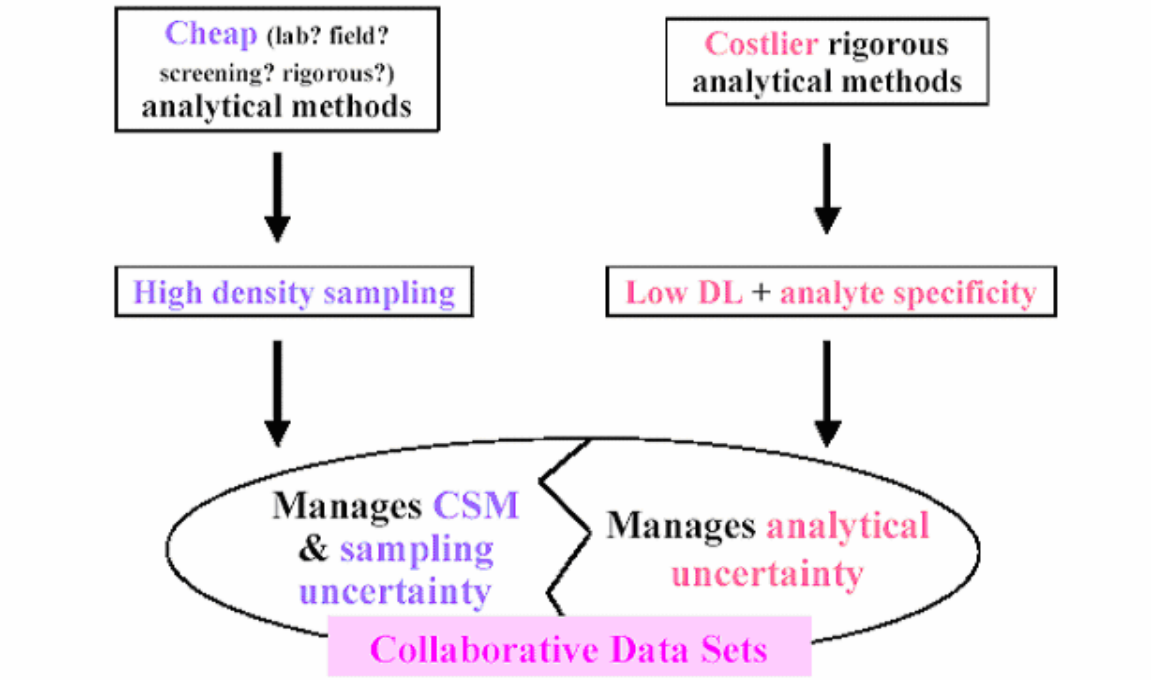

Exhibit 3. Collaborative Data Sets Increase Data Quality in Heterogeneous Matrices (DL = Detection Limit) (adapted from Crumbling et al, 2003)

highest quality, free from *any* uncertainty, and suitable and sufficient for any potential decision. A corollary of this first-generation data quality model is the mistaken impression that any data produced *outside* of traditional certified laboratories should be suspect, or what is worse – summarily rejected as unusable. But this paradigm ignores the impact of sampling uncertainty on data quality. Sole reliance on high analytical quality is insufficient to overcome the data uncertainties created by heterogeneity and sparse data points.

The first generation data quality model discourages the use of real-time tools, relegating their results to "screening" quality, while simultaneously sending the message that "screening" is never quite good enough. The label "screening" provided cover for practitioners to use poorly trained operators and inadequate documentation, further reinforcing regulators' suspicions that the data were probably untrustworthy. This self-fulfilling prophecy became a vicious circle that has hampered the environmental community's ability to learn how to use these tools efficiently and defensibly. Suppression of field methods and the failure of fixed lab analysis to provide sufficient data density led to a pattern where costly, inefficient, and flawed site characterizations were accepted as the normal state of affairs. Furthermore, the comforting illusion that fixed lab methods will robotically provide iron-clad data has long allowed the environmental community to avoid acquiring the skills it desperately needs to assess real-world data uncertainty, control for that uncertainty's impact on decision-making, and thereby improve the quality of restoration projects. As long as data users could be lulled into believing that good quality data was guaranteed by simply selecting the "approved" analytical method, the community could pay lip-service to the idea of matching data quality to data use, without ever really learning how to do it.

This is a preprint of an article published in *Remediation: The Journal of Environmental Cleanup Costs, Technologies & Techniques, Vol. 15*, No. 1, pp. 21-37. The published reprint is available through http://www.interscience.Wiley.com/

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 can only be assessed in the context of specific data use. Our experience with Triad projects has found that transitioning from the overly simplistic first-generation data quality model to the more complex (but realistic) second-generation data quality model is difficult for many regulators and practitioners. They have difficulty 1) integrating the idea of sampling uncertainty into the current data quality classification scheme, and 2) treating non-traditional analyses (often field techniques) as an equal analytical partner. To facilitate Triad projects, a new framework for integrating data sets based on data use is needed to support implementation of collaborative data sets.

To create an alternative to the traditional "analytical quality = data quality" paradigm, we have found it useful to create "data use categories" that capture the intended data use and match it to the sampling and analytical quality of the data. Unlike traditional data classification schemes, Triad-friendly data use categories cannot be defined simply according to the analytical technique used to generate the data. As noted above, the analytical technique alone is a poor predictor of data usefulness. As will be covered in much greater detail later in this paper, the same technique may produce data that falls into several different categories, even on the same project. For example, X-ray fluorescence (XRF) analysis for metals can produce data effective for supporting a wide range of decisions encompassing restricted and unrestricted land use options. The usability of XRF data depends on the analyte (XRF may report more than 10 analytes in a single analysis), method modifications, sample selection and processing, matrix interferences, and of course, the decision (e.g., contamination delineation, regulatory compliance, and risk assessment) to which the data are applied. Trying to classify data quality according to just the analytical method or laboratory certification status ignores too many important variables that greatly impact data usability.

Any proposal for a data integration framework must be able to embrace the individual factors determining data's ability to support science-based decisions. These factors arise from the intersection of the analytical technique, the regulatory oversight structure, site-specific matrix characteristics, and project goals. They include (but are not limited to):

- 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, sub-sampling, 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.

The example framework described below was originally developed to support decision-making regarding contaminated soil in Brownfields projects. The framework is offered for consideration by the cleanup community. More testing by Triad practitioners is needed to ascertain whether this scheme is general enough to work across widely different project scenarios. This paper does

This is a preprint of an article published in *Remediation: The Journal of Environmental Cleanup Costs, Technologies & Techniques, Vol. 15*, No. 1, pp. 21-37. The published reprint is available through http://www.interscience.Wiley.com/

not directly address data sets used to support designs for engineered remedial systems; it describes a data quality classification framework grounded in data use and supportive of collaborative sampling and analytical schemes able to manage data uncertainty in the context of characterization, compliance, and risk-based decisions.

DATA USE CATEGORIES DESCRIPTION

Four data use categories, briefly summarized in Table 1 on the next page and described in greater detail in the remainder of this paper, fall into two primary areas:

1) Data used primarily to rapidly develop the CSM during dynamic field programs and to manage sampling uncertainty as depicted by the left-hand side of Exhibit 3:

    a. CSM:dirty
    b. CSM:clean

2) Data used to refine ("polish") the CSM by managing analytical uncertainty relevant to regulatory and risk based decisions which require analyte-specific, analytically unbiased data sets as indicated on the right-hand side of Exhibit 3:

    a. CSM:compliance
    b. CSM:risk-calc

Regulators are most concerned with the quality of chemical data for pollutants. However, contaminant data are not the only kind of data used to develop and mature 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 project CSM. There can also be considerable overlap in how data sets are used. Cost-effectiveness and project efficiency is 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, a collaborative data strategy, that can be used to facilitate pre-field planning discussions and management of field and fixed laboratory data produced in real-time. They can be used to communicate with regulators and stakeholders about how different types of data will be integrated into a transparent, defensible, well-documented decision-making process.

This is a preprint of an article published in *Remediation: The Journal of Environmental Cleanup Costs, Technologies & Techniques, Vol. 15*, No. 1, pp. 21-37. The published reprint is available through http://www.interscience.Wiley.com/

Table 1. Summary of the Four Data Use Categories Covered in this Paper

Data Use Category (Short-Hand Designation)	Application	Activity Supported	Limitations Typical of this Category	Conditions for Use
Data Use Categories Supporting Rapid, Cost-Effective CSM Development				
CSM: dirty (used to build the CSM)	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
CSM: clean (used to build the CSM)	CSM development for situations with lower 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 requirements	Applied in conjunction with CSM: dirty using a collaborative data management strategy and dynamic work strategy
Data Use Categories Requiring More Stringent Management of Analytical Uncertainty				
CSM: compliance (used to polish the CSM)	Satisfies regulatory requirements for lab certification and strict adherence with method QA/QC and reporting deliverables	Effective for meeting expectations for demonstrating site closure and "no further action (NFA)" decisions and/or compliance monitoring	No limitations on data quality decisions with regard to analytical method; however a mature CSM is needed to establish sample representativeness	Collaborative data management strategy: samples collected after CSM has evolved sufficiently to guide selection of appropriate locations & sample processing procedures
CSM: risk-calc (used to polish the CSM)	Most stringent analytical quality from both scientific and regulatory perspective	Quantitative risk calculations to assess human health and ecological exposure from site chemicals	No limitations with strict adherence to laboratory certifications and method QA/QC; mature CSM required to establish sample representativeness	Collaborative data management strategy: sample locations identified after CSM has matured sufficiently to define exposure pathways and populations

The CSM:Dirty Data Use Category – Data that Are Effective for CSM Development in Situations With Elevated Concentrations

General Description

This data-use category includes data sets that are suitable for real-time detection, delineation, and modeling of higher concentration ("dirty") areas, especially those areas believed to exceed the established decision threshold. Using high sample throughput rates, this data use category is intended to support the kind of high-density sampling that rapidly advances the development of the CSM. Typically, the techniques used (to generate data falling into this category) report only higher analyte concentrations with confidence. At lower concentrations, results tend to be too

uncertain to support confident decisions at lower regulatory thresholds. Commonly this is because the quantitation limit and/or analyte specificity are inadequate, so lower concentration ("cleaner") areas cannot be demonstrated with sufficient confidence. Despite their lower analytical rigor, the data in this data use category will be of known quality as long as associated QC demonstrates that sampling and analysis are under control and adequate to support the intended data use. 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 another analyte(s) because a predictive relationship between them can be demonstrated (not necessarily using statistical regression). 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 non-exceedences.

To summarize, data in the CSM:dirty category 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), depending on data needs. 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 non-representative way. Increasing the number of readings to understand whether micro-heterogeneity is a problem for the matrix and analyte under investigation may control the uncertainty introduced by small sample supports. Through its contribution to CSM development, this data category is important (along with the CSM:clean category) to help establish the representativeness of other samples.

Benefits of the 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 with higher concentrations. 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 as early in the project as possible; and
- 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.

This is a preprint of an article published in *Remediation: The Journal of Environmental Cleanup Costs, Technologies & Techniques, Vol. 15*, No. 1, pp. 21-37. The published reprint is available through http://www.interscience.Wiley.com/

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.

- **Deliberate generation** of data occurs when sample processing and analytical techniques are selected for the express purpose of rapidly processing high numbers of samples. Achieving higher data density is more useful at this stage of CSM development than having high analytical quality, which would be cost-prohibitive for the needed number of samples and slow turnaround would hamper real-time decision-making. The systematic planning process determines that the type of uncertainty these data will manage in the given decision scenario can be adequately addressed with this less rigorous data set.

- **Inadvertent generation** of CSM:dirty data sets occurs when a more rigorous data set (e.g., lower detection limits) was planned, but when the results came back, data quality ended up not being as "good" as the project team had expected. Although analytical quality is not as good as planned, the data may still have some value. Although it may be inadequate for more stringent data uses (such as demonstrating regulatory compliance or calculating risk), the data may still have utility for building confidence in the CSM. In other words, data that must be rejected for a stringent data use may still be quite useful for a less stringent data use, so it need not be totally discarded. 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 (when better analytical quality was expected) include:

    1. Matrix interferences: Sample effects may degrade the performance of methods that were expected to produce more rigorous data. 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: Project planners make the dangerous assumption that using a standardized analytical method will automatically guarantee adequate data quality. But no one noticed that the method was not really appropriate for all of the target analytes or for the intended data use. 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 this project's needs. Proper planning would have determined that an alternate or modified method was needed before resources were spent generating an inadequate data set.

    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.

This is a preprint of an article published in *Remediation: The Journal of Environmental Cleanup Costs, Technologies & Techniques, Vol. 15*, No. 1, pp. 21-37. The published reprint is available through http://www.interscience.Wiley.com/

4. Instrumentation problems: Quality control may indicate that instrument, blank, or batch problems exist that unexpectedly limit the utility of data for intended purposes.

The CSM:Clean Data Use Category – Data Effective for CSM Development in Situations with Lower Concentrations

General Description

This data-use category includes data of sufficient analytical quality to delineate areas where contaminant concentrations are generally lower than the regulatory limits (i.e., "clean" or "compliant" soil). The purpose of this data use category is to identify clean areas/volumes of the site, and refine the CSM to bound "clean" areas 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 subject 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 to a data set.

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 supporting a mature CSM delineating "clean" and "dirty" areas of the site. 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 non-specific or are modifications of standard methods that have not yet 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 standard methods in order to increase sample through put. Under current laboratory certification procedures, any method modifications may be unacceptable to the certifying authority, even if the analytical performance is unchanged or even 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 to be restricted to CSM:clean and CSM:dirty uses only. As

This is a preprint of an article published in *Remediation: The Journal of Environmental Cleanup Costs, Technologies & Techniques, Vol. 15*, No. 1, pp. 21-37. The published reprint is available through http://www.interscience.Wiley.com/

the cleanup industry evolves to focus on decision uncertainty management, we hope that such restrictions will be reconsidered in the interests of promoting efficient, effective investigations and cleanups.

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 to check performance of analytical methods generating CSM:dirty data and confirm CSM:dirty data.

These data are effectively used to stratify contaminant populations for statistical purposes, or to 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). Along with the CSM:dirty category, these data are generally used as collaborative data that manage sampling uncertainties (Exhibit 3).

Determination of When Data Fall into the CSM:Clean Data Use 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 usefulness of data originally intended to serve more rigorous applications. Data can be expected to fall into the CSM:clean data use category under the following conditions:

- 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 semi-quantitatively (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 analytical uncertainty, the data are entirely suitable for supporting constrained decisions 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 lead to a high level of confidence to render a decision on cleanliness, despite some uncertainty about the actual concentration of specific analytes.

- 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 in the field or using non-traditional methods, even if the data would be considered acceptable for the intended use from a purely scientific standpoint.

This is a preprint of an article published in *Remediation: The Journal of Environmental Cleanup Costs, Technologies & Techniques, Vol. 15*, No. 1, pp. 21-37. The published reprint is available through http://www.interscience.Wiley.com/

The CSM:Compliance Data Use Category – Data Effective for Managing Analytical Uncertainty to Demonstrate Regulatory 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 (e.g., data deliverables).

Since the analytical techniques used to generate this type of data typically analyze very small sample supports (e.g., 1 to 10 grams of soil), uncertainty about the representativeness of the analytical sample may be very high. To manage this, the dynamic work strategy will call for these samples to be selected as splits from samples that had been analyzed in the field as part of CSM-building. The split samples are sent for collaborative analysis at a fixed laboratory or sophisticated mobile laboratory in order to generate the CSM:compliance data category. Split sample results help build confidence that the larger data set of field analyses is being interpreted correctly. Split samples also help to further refine or "polish" the CSM (which had been developed at a coarser scale from the CSM:dirty and CSM:clean data) with analyte-specific data and lower detection limits that can support decisions about regulatory compliance.

The fraction and selection of samples to be split for collaborative analysis should be guided by several considerations. If data use involves decision-making at an action level, a large fraction of the split samples should be focused on managing 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 entire concentration range covered by the intended regression. Either case requires knowing the approximate concentration range of the sample before selecting it for split-sample analysis. Therefore, random selection of an arbitrary percentage of samples is undesirable, because it is likely to produce a data set with a high number of non-detects or other non-informative results useless for data comparison and statistical purposes (ITRC, 2003). So 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 to be targeted with the more expensive analysis. With some advance planning on sample volume and storage, samples can be field tested and archived for later splitting as necessary. For example, non-destructive testing (such as XRF technology on homogenized samples) allows for easy archiving of the sample cups for future submission to additional laboratory analysis, while avoiding the problems created by splitting a non-homogenized sample.

Compliance data are distinguished from risk calculation data because the demands on quantitation limits, data precision and bias can be more stringent for quantitative risk calculations than for determining compliance against a regulatory threshold.

This is a preprint of an article published in *Remediation: The Journal of Environmental Cleanup Costs, Technologies & Techniques, Vol. 15*, No. 1, pp. 21-37. The published reprint is available through http://www.interscience.Wiley.com/

The CSM:Risk-Calc Data Use Category – Data Effective for Managing Analytical Uncertainty to Support Quantitative Risk Calculations

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 (to avoid a biased assessment when significant data points are non-detect) and analyte-specificity. In addition, site-wide or exposure-unit 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 from AOC driven sampling regimes may preclude site-wide (i.e., the entire site is designated as the exposure unit) risk assessments. A simple example is a site where only a portion is contaminated - say 50 percent. If the data set were comprised solely of samples taken from the AOC portion of the site, the risk assessment could be interpreted such that the risk is overestimated by a factor of two.

To minimize uncertainty in risk assessment, it is important to have low bias and good precision in the data set used to calculate contaminant exposure. Bias and precision in data are not just a function of the bias and precision of the analytical technique, but are heavily influenced by various sources of sampling-related heterogeneity. For risk assessment data, it is of utmost importance for both sampling and analytical uncertainties to be strictly controlled. Although the demands on data rigor make it desirable to obtain the best analytical quality that is technically feasible, the subsamples actually analyzed tend to be quite small. So imprecision in the data set will be high unless sampling variables are strongly managed at both macro (sampling locations) and micro (sample preparation) levels. A mature CSM that captures any significant contaminant patterning (i.e., adjacent, but distinctly different contaminant populations) and the variability within those populations should be the basis for stratifying populations, selecting sample numbers and locations, and choosing sample collection and handling procedures. Both between- (i.e., macro) and within-sample heterogeneity (i.e., micro) should be measured and controlled so that these expensive data points, produced in strict compliance with standard method QA/QC requirements, certifications and data reporting deliverables, will have maximal effectiveness for the risk-calc data use.

The more stringent data uses categories (compliance and risk-calc) are supported when the appropriate sampling and analytical techniques are selected and implemented with no confounding analytical interferences, operator error, cross-contamination, or QC problems that could cause the data to be flagged/qualified. These data sets have historically been generated in the controlled environment of a sophisticated 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 to tolerate less controlled field deployment. One such example is field-portable gas chromatography/mass spectrometry (GC/MS) instruments equipped with standardized sample preparation modules.

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 be 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,

This is a preprint of an article published in *Remediation: The Journal of Environmental Cleanup Costs, Technologies & Techniques, Vol. 15*, No. 1, pp. 21-37. The published reprint is available through http://www.interscience.Wiley.com/

produce data that is comparable across different analytical techniques, and ensure that the correct matrix population has been targeted for representative analysis (USEPA, 2003).

Since it is generally more expensive to generate data suitable for the more stringent analytical 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 less expensive ways. 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 the more expensive procedures used to create an unrestricted and unlimited land-use driven decision-making data set. This concept is illustrated in Exhibit 4, 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 was indicated in Exhibit 3.

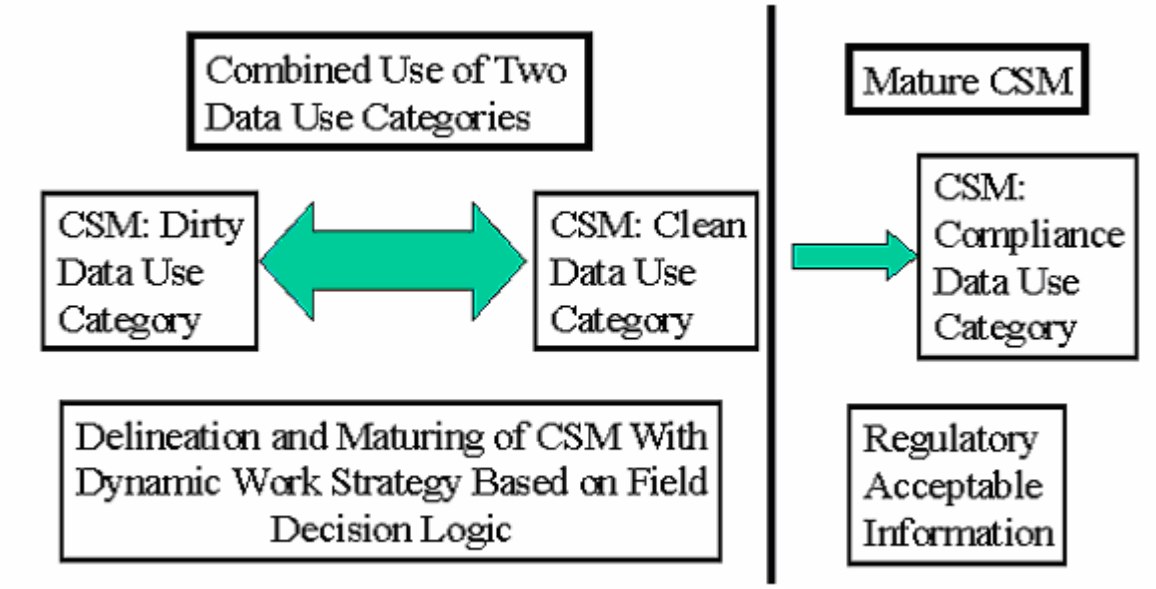

Exhibit 4. Generalized Collaborative Data Management Using Data Use Categories

SUMMARY

Judicious blending of different data-use categories maximizes the return on investment in a site characterization project because sampling uncertainties, and other important variables that could

produce misleading information, are identified and controlled. A carefully designed combination of sampling and analytical techniques manages the fundamental mismatch between the tiny volumes analyzed by traditional sampling and analysis programs 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, sampling designs can provide greater coverage, striking a balance between the need to delineate known or expected areas of contamination for more precise estimates of treatment volumes vs. the need to adequately assess the rest of the site for the presence of unanticipated contamination and to confirm areas where no further action is required.

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, the representativeness of isolated 1- or 10-gram analytical samples is unknown. When interpreting data results, the data user does well to ask, 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 hotspots. 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. This paper proposes a framework to structure data quality in a way that supports scientifically defensible decisions and efficient projects, while also being used to meet regulatory oversight objectives.

REFERENCES

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

ASTM. (1998). Standard Guide for Accelerated Site Characterization for Confirmed or Suspected Petroleum Releases. American Society of Testing Materials. ASTM E-1912-98. West Conshohocken, PA.

ASTM. (1998a). Standard Practice for Expedited Site Characterization of Vadose Zone and Ground Water Contamination at Hazardous Waste Contaminated Sites. American Society of Testing Materials. ASTM D-6235-98. West Conshohocken, PA.

This is a preprint of an article published in *Remediation: The Journal of Environmental Cleanup Costs, Technologies & Techniques, Vol. 15*, No. 1, pp. 21-37. The published reprint is available through http://www.interscience.Wiley.com/

ASTM. (1999). Guide for Developing and Implementing Short Term Measures or Early Actions for Site Remediation. American Society of Testing Materials. ASTM D-5745-95. West Conshohocken, PA.

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

CTLUSTP. (2000). Expedited Site Assessment: The CD (Version 1.0) – UST Site Investigation Guidance for a New Millennium. Connecticut Department of Environmental Protection, Leaking Underground Storage Tank Program. Hartford CT.

Crumbling, D, C. Groenjes, B. Lesnik, K. Lynch, J. Shockley, J. VanEe, R. Howe, L. Keith, G. McKenna (2001). Applying the Concept of Effective Data to Contaminated Sites Could Reduce Costs and Improve Cleanups. *Environmental Science and Technology*, 35(19): 405A-409A. October 1, 2001. Available at http://cluin.org/download/char/oct01est.pdf

Crumbling, D.C., Griffith, J., Powell, D.M. (2003). Improving Decision Quality: Making the Case for Adopting Next-Generation Site Characterization Practices. *Remediation* 13(2): 91-111 (Spring 2003). Available at http://cluin.org/download/char/spring2003v13n2p91.pdf

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 (Spring 2004).

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. Available: 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 (Autumn 2003).

Robbat, A. (1994). 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," for the U.S. Army Environmental Engineering Center, Aberdeen, MD, May 1994. Tufts University. Boston MA.

Robbat, A. (1997). Dynamic Workplans and Field Analytics: The Keys to Cost Effective Site Characterization and Cleanup. Medford, Massachusetts: Tufts University Center for Field Analytical Studies and Technology. Boston, MA. Video and transcript available at http://cluin.org/studio/video.cfm

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

This is a preprint of an article published in *Remediation: The Journal of Environmental Cleanup Costs, Technologies & Techniques, Vol. 15*, No. 1, pp. 21-37. The published reprint is available through http://www.interscience.Wiley.com/

USEPA. (1997). Expedited Site Assessment Tools For Underground Storage Tank Sites: A Guide For Regulators, EPA 510-B-97-001, Office of Underground Storage Tanks, Office of Solid Waste and Emergency Response, Washington, DC. Available http://www.epa.gov/OUST/pubs/sam.htm

USEPA. (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. United States Environmental Protection Agency, Washington, DC. Available at http://cluin.org/download/char/treefruit/wtfrec.pdf

USEPA. (2003). Guidance for Obtaining Representative Laboratory Analytical Subsamples from Particulate Laboratory Samples, EPA/600/R-03/02. United States Environmental Protection Agency, Washington, DC. Available at http://www.epa.gov/esd/tsc/images/particulate.pdf

USEPA. (2004). Improving Sampling, Analysis, and Data Management for Site Investigations and Cleanups, Office of Solid Waste and Emergency Response, EPA-542-F-04-001a. United States Environmental Protection Agency. Washington, DC. Available at http://www.cluin.org/download/char/2004triadfactsheeta.pdf

Woll, B., Mack, J., Ellerbusch, F., Vetter. J. (2003). Facilitating Brownfields Transactions Using Triad and Environmental Insurance. *Remediation* 13(2): 113-130 (Spring 2003). Preprint available at http://www.triadcentral.org/ref/doc/Remediation_preprint_Triad-Insurance.pdf

ABOUT THE AUTHORS

James Mack, York Center for Environmental Engineering and Science, New Jersey Institute of Technology, Newark, New Jersey 07102-1982; phone: 973-596- 5857; fax: 973-229-2415; mack@adm.njit.edu

James Mack is a director at the Northeast Hazardous Substances Research Center where he provides technical assistance to cleanup programs and projects. He holds a B.S. in Geology from Waynesburg College and a M.S. in Earth Science from Adelphe University.

Deana M. Crumbling, USEPA Office of Superfund Remediation and Technology Innovation, 1200 Penna. Ave., NW, Washington, DC 20460; phone: 703-603-0643; fax: 703-603-9135; crumbling.deana@epa.gov

Deana Crumbling has worked in the hazardous waste site cleanup arena over the past 12 years, and has been at USEPA since 1997. She is an analytical chemist with clinical, industrial, and research experience. She holds a B.S. in Biochemistry, a B.A. in Psychology, and an M.S. in Environmental Science.

Fred Ellerbusch, York Center for Environmental Engineering and Science, New Jersey Institute of Technology, Newark, New Jersey 07102-1982; phone: 973-596-6341; ellerbus@adm.njit.edu

Fred Ellerbusch, P.E., DEE, is a director at the Northeast Hazardous Substances Research Center and a faculty member at the University if Medicine and Dentistry of New Jersey (UMDNJ) School of Public Health. He is a Ph.D. candidate and holds a MPH from UMDNJ as well as a M.S. and B.S. in Engineering from the NJIT.