GEOTECHNICAL SPECIAL PUBLICATION NO. 179

GEOCONGRESS 2008

CHARACTERIZATION, MONITORING, AND MODELING OF GEOSYSTEMS

PROCEEDINGS OF SESSIONS OF GEOCONGRESS 2008

March 9–12, 2008
New Orleans, Louisiana

SPONSORED BY
The Geo-Institute of the American Society of Civil Engineers

EDITED BY
Akram N. Alshawabkeh
Krishna R. Reddy
Milind V. Khire

Published by the American Society of Civil Engineers

Six Sigma Approach to Sustainable Institutional Environmental Data Management

Christopher M. French[1], Neno Duplancic[2], Greg Buckle[2], Marian Carr[2], Josh Jaffe[2], Jennifer Holland[3], Dak Patel[4], William Colby-George[5], Rene Surgi[6], Laura Drachenberg[6], Tom Conklin[7] and Chuck Sharpe[7]

[1]Honeywell International Inc., Morristown, NJ.; Chris.French@Honeywell.com
[2]Locus Technologies, Inc., Mountain View, CA.
[3]CH2M HILL, Phoenix, AZ.
[4]CH2M HILL, Philadelphia, PA.
[5]Mactec, Portland, OR.
[6]AESI, Glencoe, IL.
[7]Parsons Inc., Syracuse, NY.
[8]O'Brien & Gere Inc., Syracuse, NY.

ABSTRACT: The results of a 3-year six sigma evaluation of a centralized corporate remediation data management system are presented. The primary focus of the study is to improve electronic management of remediation data generated for the corporate environmental remediation function. The examination is unique in that no prior body of work has applied six sigma methods to environmental remediation data management. Both qualitative and quantitative six sigma tools have been applied in the study. Metrics are presented illustrating significant improvements in cost, quality, and cycle time since implementation of the system. A cost function is derived to predict normalized costs for data management as a function of the number of records in a database based upon a statistical population of 110 remediation sites and over 11 million records. The importance of remediation data management is examined within the context of process sustainability from the standpoint of protection of human health and environment, improved regulatory compliance, and greater transparency. The study is relevant to the state of environmental remediation within the context of more stringent enforcement through the regulatory agencies and the courts, an intensifying complexity of state and federal electronic data delivery (EDD) requirements, a ratcheting downward of cleanup standards, lower analytical detection levels, increasing requirements for capture and retention of analytical metadata, continued reliance on containment and institutional controls, and a parallel increasing demand for data that quantifies the nature, extent, and temporal variability of contamination. Application of six sigma metrics results in more-effective institutional

stewardship manifested by reduced cycle time, significantly reduced cost, and enhanced data quality and defensibility through the long-term remediation lifecycle, which can span decades. A case study is presented for a complex, multimillion-dollar site remediation effort.

INTRODUCTION AND PURPOSE

The study examines the application of tools and strategies, collectively referred to as six sigma methods, to improve the process for electronic management of analytical laboratory data generated during the execution of environmental remediation projects in the industrial sector for a multinational industrial conglomerate. Both qualitative and quantitative tools can be used. Qualitative tools assist in scoping and defining process variability and resultant process defects and in the design of controls to address the defective process. Quantitative six sigma tools measure how effectively defects and variations are eliminated from products, processes, and services. Six sigma statistics, where quantifiable, provide metrics to measure process defects and improvements garnered through implementation of a control plan. A six sigma process evaluation comprises five steps: (i) process definition, (ii) measurement, (iii) analysis, (iv) process improvement, and (v) establishment of process controls.

Environmental Remediation Data Management

In general, the site remediation process may be broken down into three major phases: (i) site characterization, (ii) remedy implementation, and (iii) operations and maintenance (O&M). Project analytical data are generated in the course of all project phases. Analytical data uses over the lifecycle of the site cleanup may span years or decades. Many stakeholders will access the same data for disparate uses. In short, for the site process to be sustainable, the site institutional knowledge must be accessible over many years and project phases, and future users must be satisfied that the historical ("legacy") data are usable, and their quality can be assured.

Problem Statement

Electronic data delivery (EDD); data checking, verification, and validation; and database administration can be very complex undertakings. Standardizing these processes to the extent practicable is key to sustainable analytical data management. The industrial sector is generally lagging behind the government sector in developing standardized approaches to managing environmental remediation data. Whereas a variety of federal (USEPA, USDOD, USDOE) and state agencies have adopted uniform formats for managing their site remediation data and have imposed regulatory requirements for data submittals by responsible parties undertaking private sector remediation efforts, environmental consulting businesses and the industrial client base they serve are only now beginning to seriously address the data management process at the level of the enterprise (as opposed to the specific project). This examination is prompted largely in reaction to the myriad electronic regulatory reporting requirements that have emerged in the past few years, in response to the burgeoning cost and complexity of remediation projects, and in recognition of the protracted cycle time involved in project execution and the strategic necessity of implementing a long-term risk management strategy.

Major business drivers include the following:

- Ensuring process sustainability from the standpoint of protection of human health and environment, improved regulatory compliance, and greater transparency.
- Managing and sustaining liability and risk over a protracted project and regulatory cycle characterized by more stringent enforcement through the regulatory agencies and the courts, an intensifying complexity of state and federal EDD requirements, a ratcheting downward of cleanup standards, lower analytical detection levels, increasing requirements for capture and retention of analytical metadata, continued reliance on containment and institutional controls, and a parallel increasing demand for data that quantify the nature, extent, and temporal variability of contamination.
- Managing the significant cost and cycle time challenges to find, track, and retrieve information.
- Recognizing the increased reliance on consultants to provide information and data, thus driving costs, reducing direct access to data, and potentially compromising the long-term sustainability of the process.
- Acknowledging that the perceived absence of reliable legacy data creates redundant work or project rework over the site project lifecycle, which spans years or decades.
- Losing opportunities to leverage the considerable investment in information over the duration of the site lifecycle or across the enterprise ("data mining").

Corporate environmental remediation departments and their counterparts in regulatory agencies are changing as their processes are increasingly driven by economics, regulatory compliance, risk management, and sustainable corporate governance practices. A great deal of an entity's success in managing a remedial program will depend on how technical data are managed. The premise is that standardizing and digitizing the analytical data management process will result in fewer defects in the process and more systematic attainment of data quality objectives at reduced cost and time. These factors profoundly influence the sustainability of the corporate function.

APPLICATION AND RESULTS

A pareto analysis of spending for the two major cost categories of remediation (investigation and remedial-action-plus-O&M) identified 45 projects that accounted for roughly 80% of total corporate spending. The analysis indicates there is an 80% overlap between the 30 largest projects by analytical cost and 15 largest projects by overall cost. Twelve of the 15 largest projects by overall cost account for 65% of overall laboratory analytical spending. In addition, analytical costs tend to be evenly split between the remedial investigation phase and remedial-action-plus-O&M phases. The financial analysis was combined with laboratory program information on each of the high-cost projects. This was to ensure that the baseline exercise reviewed projects with a reasonable distribution of analytical methodology.

"Data management cost" was defined as that cost incurred for entering, checking, and verifying data and for entering validation qualifiers in an electronic repository. It was concluded that a normalized unit cost, data management cost per record (independent of project validation requirements, specifics of the chosen data repository, and data management process invoked) was the best measure of efficiency for the purposes of the analysis.

Process maps were developed for three benchmark projects. Combined with the cost-per-record information, these annotated process maps led to the following conclusions about reasons for differences in cost per record and for high per-record costs in general:

- Validation cost impact on the data entry process was significant because the validation in each instance was a highly manual process.
- A distinction was noted in costs related to whether the data management system stores all chemistry data and metadata in its repository or only primary chemistry. Data entry costs can easily double if QC data are stored, particularly for non-automated processes.
- Automation of the data import process with imbedded data-checking routines significantly improves process efficiency, reduces cycle time, and ensures data integrity.

Chemistry data from one project was examined for data defect rates, where "defect rate" is defined as percent of unusable data. The data were determined to have a 38% defect rate.

A temporal process map was developed to examine the evolution of factors that affect sustainability (over several project phases) attributable to problematic "legacy data" for a typical site or project for which standards for data management have not been developed and enforced from project outset. The analysis demonstrated that failure to incorporate mission-critical legacy data prior to the early critical project milestones (pre-remedy selection) can substantially impact the sustainability of the long-term project.

Stakeholder surveys were developed for 32 projects. The survey of corporate remediation managers indicated that the overall data management process was poorly understood because this function was typically undertaken by consultants with no direct involvement of corporate management. For corporate projects surveyed, 32% used no data management system, 23% used a desktop commercial software program, and 27% used a proprietary system. The types of data management system used for the remaining projects were unknown but may be assumed to be in the form of a desktop spreadsheet application. Of vendor responses, 47% indicated they used a stand-alone, custom-built database system, and 31% of respondents used a commercial system. Two respondents (6%) indicated they used a proprietary system, and 17% used no database management system. Specifications for a database had been developed in 54% of the projects surveyed, and 67% of respondents reported employing a standardized schema for the project. Of the database systems employed, 44% resided in Excel spreadsheets and 37% in Access database structures. In general, from the standpoint of sustainability, it was determined that long-term access to the data by the institution was potentially compromised.

Vendors were asked to describe what type of data entry was performed (automated, manual, or a hybrid process) and what type of data is stored (analytical only or analytical data plus quality control (QC) data). Roughly one-half of project data management systems may be classified as "hybrid," meaning that a combination of manual and automated processes is used for data entry. One-third may be classified as fairly automated systems. Manual data import routines composed 13% of projects.

A side-by-side comparison of data storage practices (analytical data only versus analytical and QC data) and the prevalence of imbedded procedures to perform automated data integrity checks was evaluated. Automated systems show a slightly higher tendency than hybrid systems to store both analytical and QC data and to incorporate embedded routines that check data integrity and data QC relationships. Of the four projects using manual data entry routines, three store both analytical and QC data, but none incorporates routines to enforce data integrity and QC relationships.

Automated systems present a higher incidence of data checking and verification routines overall, with fairly substantial margins reported for procedures designed to check field format and required fields, valid value compliance, comparison to sample tracking tables and normal and QC data-batching checks. Field-formatting checks, location/sample I.D. checks, and duplicate EDD value checks were the only three routines that are performed in over 50% of the cases.

Data verification and validation activities are important from the standpoint of risk mitigation and cost. The majority of management respondents (72%) indicated that some work would be required to defend the quality of their remediation data, as opposed to data quality being readily defensible (17%) or defensible only with substantial research and effort (11%).

Management and suppliers were surveyed to assess federal and/or state agency reporting requirements. The management survey indicated that approximately 56% of respondents believe specific electronic data formats were required by the lead regulatory agency at the time the survey was undertaken, in early 2002. A follow-up survey undertaken in 2006 indicated that electronic standards for data reporting were required or in development in a significant majority of cases examined.

Vendors reported their companies' primary methods for data archiving as paper (34%) or backup tapes (31%), followed by CD (22%) and floppy disk (13%). However, when asked how data are archived on behalf of the corporation, the vendors reported by floppy disk (38%), CD (21%), backup tapes (20%), hard copy (16%), or none (5%). This result, when viewed in concert with the proprietary or customized nature of most database applications utilized, suggests that capturing mission-critical legacy data would be necessary to ensure the sustainability of the site management process.

Detailed cost and record data are available for 19 projects based upon information provided in the vendor survey. Automated systems are most prevalent in the dataset (10), followed by hybrid (6) and manual (3). The automated systems tend to contain a larger number of records than either the hybrid or manual systems. In other words, larger data sets command the use of more highly automated systems.

A cost function was derived for normalized cost (cost per record) as a function of log ($N_{records}$) for manual, hybrid, and automated systems (Figure 1). Costs for automated data entry systems were juxtaposed against those for manual and hybrid data entry systems. Both cost functions appear to represent a gamma function. As the number of data records in the project database increases, the normalized cost per record decreases. The hybrid/manual gamma function presents a steep decay function. On the other hand, the slope of the cost decay function for automated systems is comparatively low, and the average value for data management is lower. The gamma function is similar to cost decay curves observed for industrial mass production processes.

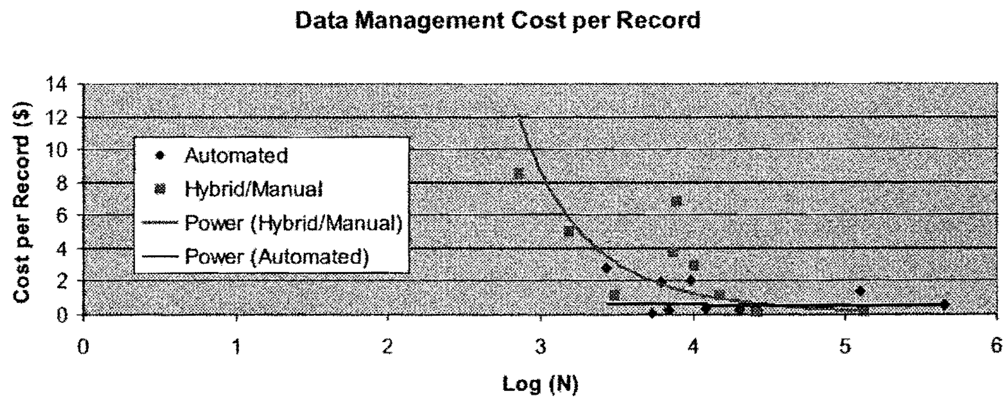

FIG. 1. Normalized cost per record for automated versus manual/hybrid database systems.

Median values for automated, hybrid, and manual unit costs per record were $0.47, $1.18, and $3.84, not including data validation costs. Statistical evaluation of the cost data shows a significant difference between the automated and manual unit costs ($p < 0.05$) but confusion between automated and hybrid systems, and hybrid and manual systems, most likely due to the wide variance in hybrid system unit costs illustrated in the statistical comparison (Figure 2).

Variability in a process is a major factor in six sigma evaluations and an important contributor to sustainable processes. A statistical evaluation of variance indicates that significant differences exist between the data sets. It is concluded that the hybrid/manual approach introduces more variability into the cost structure.

Case Studies—Proof of Concept

Two case studies were completed to evaluate cycle time, cost, and quality metrics associated with process automation and standardization. The first case study involved examining monthly analysis and reporting under the National Pollutant Discharge Elimination System (NPDES) program. This bench-scale study quantified cycle time improvements for a time-critical process. The second bench-scale study quantified cost and quality improvements associated with a quarterly monitoring program for an intensive post-remedy monitoring program at a site subject to compliance with a federal consent decree. For the NPDES study, minimum cycle time improved 30%, and maximum cycle time improved 51%. Costs for the quarterly monitoring program, as measured for reporting and validation, have been reduced 62% and 67%, respectively. For the second case study, the earlier benchmarking exercise had identified a 38% defect rate for the monitoring project data. Data defects were completely eliminated. In the case of the monitoring program, additional savings of

$90,000 per year were realized once data trends could be observed and reductions in the monitoring regimen were made.

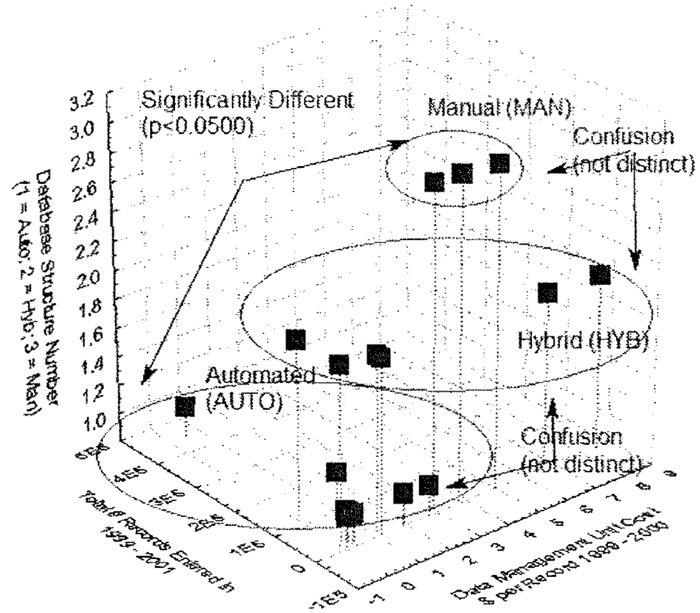

FIG. 2. Three-dimensional surface plot of normalized cost per record.

Additional case studies involving examination of cycle time and costs savings associated with the data management program are highlighted below:

- For a major site investigation, a $3,000 investment in up-front field data automation generated $100,000 in savings.
- Standardization of chain-of-custody (COC) and electronic data reporting resulted in savings of 5% in data management time for COC management, time savings of between 0.5 and 1.0 hours per EDD, and a reduction of time of 20 hours per report draft.
- Automation resulted in a 50% savings on a data validation effort ($50,000) and a 70% savings on reporting associated with a data intensive investigation effort subject to strict regulatory requirements.

Process Improvement and Control

A six sigma process evaluation incorporates five fundamental steps: (i) define, (ii) measure, (iii) analyze, (iv) improve, and (v) control. Following evaluation of the first three steps summarized in the preceding section, a process improvement and control plan was implemented to automate and standardize the corporate remediation data management function. A pilot program was implemented for three large, data-intensive projects including a complex, multimillion-dollar site remediation effort using a centralized, web-based data management system that included retention of analytical, geochemical, geotechnical, hydrological, geological, and mineralogical data. At the successful conclusion of the pilot program, the data management system was adopted across the entire remediation portfolio. At present, the system comprises 115 sites with 11 million records. Important factors contributing to the success of the program include (i) management's mandate for uniform application of the process across the remediation portfolio by contract engineering firms and laboratories, (ii) systematic, centralized management of the process change using a core team of experts aligned with regional centers of excellence, (iii) phased implementation to monitor and manage process improvements and (iv) tracking of metrics to sustain the implementation cycle and communicate the continuing success of the program to management.

"Before" and "after" cost, cycle time, and quality improvements have been tracked since implementation of the system. A cost function has been derived to predict normalized costs for data management. It is estimated that the process improvements have resulted in annual savings of $1.2 million, based upon statistical evaluation of cost data for the first 3 years of full implementation. Cycle time metrics for analytical laboratories and data management contractors were tracked through an automated reporting process, resulting in significant improvement in analytical turnaround time and completion of data reporting, verification, and validation. Laboratory analytical data quality metrics (blank, laboratory control sample, matrix spike, surrogate recovery) were also reported in the form of a management flash report used for ranking of laboratory performance.

SUMMARY AND CONCLUSIONS

The results of a 3-year six sigma evaluation of a centralized corporate remediation data management system are presented. The primary focus of the study was to improve electronic management of data generated for the corporate environmental remediation portfolio. The importance of remediation data management is presented in the form of a problem statement that examines process sustainability from the standpoint of protection of human health and environment, improved regulatory compliance, and greater transparency. The study is relevant to the state of environmental remediation within the context of more stringent enforcement through the regulatory agencies and the courts, an intensifying complexity of state and federal EDD requirements, a ratcheting downward of cleanup standards, lower analytical detection levels, increasing requirements for capture and retention of analytical metadata, continued reliance on containment and institutional controls, and a parallel increasing demand for data that quantifies the nature, extent, and temporal variability of contamination. Both qualitative and quantitative six sigma tools were applied in the study. Metrics indicate significant improvements in cost, quality, and cycle time since implementation of the system. Application of six sigma metrics results in more effective institutional stewardship manifested by reduced cycle time, significantly reduced cost, and enhanced data quality and defensibility through the long-term remediation lifecycle. Indirect improvements include enhanced accessibility, portability and collaboration, the ability to leverage the knowledge management and data-mining functions through a centralized repository, informed decision making for risk management, and enhanced corporate regulatory compliance.