INTERSTATE TECHNOLOGY & REGULATORY COUNCIL

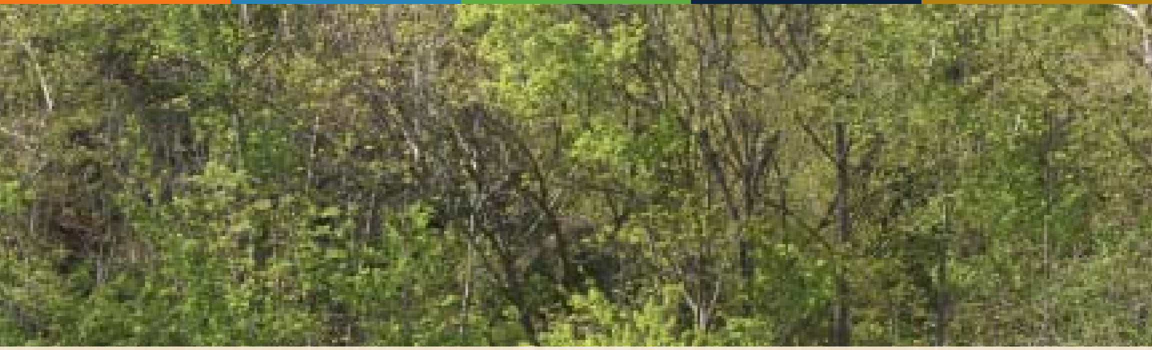
Warning! This document has not been amended since publication. Some content may be out of date and may no longer apply.

INTERSTATE TECHNOLOGY & REGULATORY COUNCIL

Technology Overview

Data Management, Analysis, and Visualization Techniques

Fourth in a Series of Remediation Process Optimization Advanced Topics



March 2006

Prepared by
The Interstate Technology & Regulatory Council
Remediation Process Optimization Team

Technology Overview

DATA MANAGEMENT, ANALYSIS, AND VISUALIZATION TECHNIQUES

Fourth in a Series of RPO Advanced Topics

March 2006

Prepared by
The Interstate Technology & Regulatory Council
Remediation Process Optimization Team

50 F Street NW, Suite 350, Washington, DC 20001

ITRC (Interstate Technology & Regulatory Council). 2006. *Data Management, Analysis, and Visualization Techniques*. RPO-5. Washington, D.C.: Interstate Technology & Regulatory Council, Remediation Process Optimization Team. www.itrcweb.org.

Disclaimer

Acknowledgments

The members of the Interstate Technology and Regulatory Council (ITRC) Remediation Process Optimization (RPO) Team wish to acknowledge the individuals, organizations, and agencies that contributed to the 5-part series on advanced RPO topics. The following individuals from state and federal agencies, and the private sector are active members of the RPO Team and supported the preparation of these documents:

- New Jersey Department of Environmental Protection–Tom O'Neill–Co-team Leader
- South Carolina Department of Health & Environmental Control–Sriram Madabhushi–Co-team Leader
- California Department of Toxic Substances Control–Ning-Wu Chang
- Florida Department of Environmental Protection–Bheem Kothur
- Georgia Department of Natural Resources–Christopher Hurst
- South Dakota Petroleum Release Compensation Fund–John McVey
- Virginia Department of Environmental Quality–Tom Modena

- U.S. Air Force–Don Gronstal, Rod Whitten, Javier Santillan
- U.S. Army Corps of Engineers–Dave Becker
- U.S. Navy–Karla Harre
- U.S. Department of Energy–Beth Moore
- U.S. Environmental Protection Agency–Kathy Yager, Richard Hammond, Pamela Baxter, Ellen Rubin
- Lawrence Livermore National Lab–Maureen Ridley

- Battelle Corporation–Russell Sirabian
- Booz Allen & Hamilton–Joann Socash
- Dajak, LLC–Mark Kluger
- Intergraph Corporation–Tanwir Chaudhry
- Mitretek Systems–John Horin, Patricia Reyes
- Northeastern University–Mary J. Ondrechen
- Remedial Operation Group, Inc.–Bud Johnson
- S.S. Papadopoulos and Associates, Inc–Michael T. Rafferty, P.E.
- SRS/Westinghouse–Kevin Brewer

Special thanks goes to the primary authors of this document on Data Visualization: Richard Hammond, USEPA, Maureen Ridley, LLNL and Sriram Madabhushi, SCDHEC.

Data Management, Analysis, and Visualization Techniques

"The most expensive sample ever collected is the one whose results are never used."

Patricia Ottesen,
Lawrence Livermore National Laboratory Data Administrator

Introduction

This overview introduces the reader to the basic concepts of Data Management, Analysis and Visualization Techniques (DMAVT). In 2004, the Interstate Technology and Regulatory Council (ITRC) Remediation Process Optimization (RPO) Team developed a technical regulatory guidance document titled, *Remediation Process Optimization: Identifying Opportunities for Enhanced and More Efficient Site Remediation*. Based on feedback to the RPO training and continued research into the topic, the RPO team identified the need for detailed information on DMAVT. Data Management, Analysis, and Visualization Techniques in some ways are important tools in successfully measuring the progress of a remediation or a monitoring program. This overview will further develop the basic concepts of DMAVT and their potential application to site rehabilitation projects.

Figure 1. A common means of data storage even in 2005 in many agencies.

Why is there a need for DMAVT? Many project managers are not familiar with these techniques. Several of these reports are stored in file rooms and stacked in hallways as shown in Figure 1. and not much information is extracted out of them. Vast amounts of data are collected at a typical cleanup site and submitted to regulatory agencies in thick reports. Over a period of years, these data accumulate and become unmanageable, without any decisions toward project management being made. In many cases not many useful inferences are drawn from these data. It is also common to see that these data are not converted into a form that can be easily presented to stakeholders in order to assess the progress (or lack there of) made toward achieving cleanup goals at a site. This document summarizes useful techniques to manage data gathered in the field including existing data available to project managers, techniques to extract useful information out of those data, be able to analyze data, and make important decisions through appropriate data visualization and presentation techniques.

Who We Are and the Intended Audience

The ITRC is a state-led coalition of regulators, industry experts, citizen stakeholders, academia, and federal partners that work to achieve regulatory acceptance of innovative environmental technologies. This coalition consists of 46 states and a network of some 7,500 people who work to break down barriers, reduce compliance costs, and make it easier to apply new technologies to solve environmental problems. ITRC helps maximize efficient use of state resources by creating a forum where innovative technology and process issues are explored. Together, the team members are building the environmental community's ability to expedite quality decision making while protecting human health and the environment.

This overview has the following intended audience who are involved in either remediation process (RPO) or PBM of hazardous site remediation projects:

- State and federal regulators
- Facility owners and operators
- Engineers and consultants
- Interested stakeholders

States and federal agencies play multiple roles in the RPO and PBM processes: as regulators, and as facility owners and operators when public funds are used to conduct site remediation work. As regulators, state and federal agencies are charged with protecting human health and the environment. Also, facility owners, private or public, have the greatest interest in achieving the goals of the specific site remediation project. In addition, the engineering and consulting community who guide and provide professional opinions to the owners must have a deep working knowledge of techniques that can ensure fast and effective site remediation. To understand PBM and be full participants in environmental remediation efforts, public stakeholders must not only understand technologies used at sites, but also the underlying technical basis that supports the decision-making process. This document is intended as an introduction to the DMAVT, however, users are encouraged to refer to the references provided at the end of the overview for additional information.

This overview is part of a five booklet series: *Performance-based Management, Analysis of Above Ground Treatment Technologies, Exit Strategy Analysis, Data Management, Analysis, and Visualization Techniques*, and *Life Cycle Cost Analysis*; each is an excellent resource for moving forward on their RPO and PBM projects

Why a need to understand Data Management, Analysis, and Visualization Techniques?

As the technology associated with data management has evolved, terms that were appropriate in the past have also evolved to take on more specific meanings. Prior to the powerful visualization tools that are now available at the desktop PC level, the term Data Management (DM) was sufficient to cover much of the work conducted under what may now be more appropriately seen as data management - the physical movement and control of data. Data visualization (DV) provides means to better present the data gathered at a site in order to effectively communicate the processes that are taking place at a site as well as provide visualization of the outcomes of various proposed remedies to broad groups of stakeholders. Charting, graphing, contour plots, etc are examples of visualization tools. Data Analysis (DA) on the other hand is the process through which meaningful inferences are made in order to develop conclusions from these data. These DMAVT of course consider into account the experience levels of various intended end user audiences and the needs of organizations involved in data management. Detailed definitions and discussions of the concepts are discussed below.

Data From the Field

Most parties involved in investigations and remediation at contaminated sites have Quality Management Plans for their organizations and Quality Assurance Project plans for detailing the Quality Assurance and Quality Control (QA/QC) program for individual sites. These QA/QC programs are essential for generating data of known and defensible quality. Each aspect of the environmental monitoring program, from sample collection to data management must address and meet applicable quality standards.

Field sampling QA involves many practices that minimize error and evaluate sample performance, including using standard operating procedures, chain-of-custody procedures, instrument cleaning and calibration, and use of QC samples. Laboratories also have extensive, written QA/QC programs including instrument calibration procedures, analysis of duplicate, blank and spiked samples, and this QA/QC data is usually reported with the monitoring results. Verification and validation of field and analytical data collected for environmental monitoring and restoration programs are necessary to ensure that data conform with applicable regulatory requirements. Validation of field and analytical data is a technical review performed to compare data with established quality criteria to ensure that data are adequate for intended use. The extent of project data verification and validation activities is based upon project-specific requirements.

Electronic Data

Electronic data are an important and integral part of any environmental Data Management, Analysis, and Visualization (DMAV) systems. Although implementation and maintenance of these systems can appear costly and require specialized resources and skills, the benefits of increased abilities for analysis and communication are significant. The term “electronic data” incorporates both the data stored and manipulated as part of the tool system, and the transmittal of data to interested parties (e.g., regulatory agencies). The pros of electronic data include:

- Easier analysis of large datasets
- Ability for more accurate and up-to-date information
- Increased ability to combine data from multiple projects

On the other hand, the primary drawbacks to electronic data include:

- Change from previous practices
- Required information management procedures and standards that require specialized skills
- Agreement on reference codes and standards to facilitate interoperability (i.e., data dictionary and entity relationships)

Currently, a number of formats for electronic environmental data exist (EPA 2005) and agreements between the state, federal, and owners/projects need to be achieved. Also, implementation and/or pilot testing of new systems are not insignificant and require significant planning and resources.

Using the Internet to transmit and access data is preferred as it facilitates standardization of formats/views and allows timely data uploading, analysis, and retrieval. However, special considerations may be necessary for large projects where transmittal of data on CDs may be more appropriate. Access of data via the Internet needs to also consider security issues to not allow changes without proper authorization, and viewing restrictions may also be appropriate in some cases for proprietary and legal reasons. These issues should be worked-out between the projects and regulatory agencies with clear roles and responsibilities. Data validation requirements should also be crafted, in particular to define what data should be used for decision making. On this line, accommodation for different qualities of data and different data usage should be made.

Data Management

Data Management is a set of processes and procedures that an organization puts in place to ensure that its data are providing their maximum usefulness to the organization. The activities include:

- Strategic data planning
- Data element standardization
- Information management control
- Data synchronization
- Data sharing
- Database development

According to the West European Road Directors Guide for Data management,

> "Data management at its simplest level is purely good housekeeping - ensuring the data you want are accessible when you want them, and provided at a cost and quality that meets your needs. … Most importantly, data management is about understanding data - turning data into useful information." (WERD 2003)

Also, data management system is a computer application used to input, store, process, analyze, and output data.

What are the benefits of good data management?

The benefits of good data management are reflected through: more efficient and cost effective decision-making, maximizing use of the data, avoiding duplication, integration and interoperability, improving access and communication, and facilitating collaborations. Organizations are accountable for their decisions, and good data management provides the necessary audit path to demonstrate the basis for their decision-making.

As part of a remediation process optimization (especially the monitoring optimization) review, it is often an opportunity to evaluate an existing cleanup system. In many cases, evaluation of gathered data would be possible using a DMVAT and suggesting recommendations on the effectiveness of remediation systems. For a new system that is put in place, a good DMAVT can make tracking the progress toward cleanup an easier task. This will in turn help in effective and efficient cleanups at contaminated sites. For example, a series of plume isoconcentration maps over a period of time can provide valuable insights into the progress towards cleanup goals at a contaminated site.

Traditional data management practices used to be confined to data gathering and storage. Most often, these were just a data storage or warehouse kind of systems. Nowadays, advances in computer data storage and retrieval have made data more accessible to project managers and to make meaningful use of such data. Simplified tools (e.g., an excel spreadsheet or an access database) to sophisticated packages (GIS tools) can be used to manage and manipulate data, to visualize from an objective perspective, and to make more appropriate decisions about the remediation processes.

Data Analysis

Data analysis can be loosely divided into *classical* and *probabilistic* statistical models. Classical statistical models work well when the object of the model is well defined (for example, when playing cards, one would never expect to draw a 3 of clubs). Since the beginning of the 20th century, a different form of statistics, probabilistic statistics, has had a growing impact for modeling less regular systems such as environmental systems. It is not hard to imagine that the selection of an underlying statistical model will have strong bearing on the type of data required for collection. As well, the visualizations that are possible from various types of data collection schemes can vary significantly.

Another set of terms that is frequently used in conjunction with statistical models is *analytical* versus *numeric*. Although the use of these two sets of terms (classical versus probabilistic and analytical versus numeric) are not identical, often classical models use analytical analysis (solving a single, linear equation such as Darcy's flow equation) while probabilistic models use numeric analysis (such as Monte Carol simulation).

Data Visualization

"Data" always has the problem of providing only a subset of the whole. In a sense then, even the data contained in a database or a statistical analysis is a form of visualization. However, in this document, the ITRC RPO team considers visualization to be a further step removed from the data. The person or people responsible for creating visualizations carry an extra responsibility because the visualization must try to convey an accurate, precise and representative view of the data, while making that visualization sufficiently refined for the intended end user.

Data visualization generally falls into three categories – *charts, graphs,* and *multimedia*. Charts are perhaps the oldest type of data visualization. Charts include products such as bar charts, histograms, scatter plots, box and whisker plots and a host of other chart types. Some samples are generated from visualization tools are shown in Figure 2. Their basic objective is to compare the rate of change between one value and another, for example the price of gasoline per month or the temperature per hour.

Graphs, including what have been traditionally called maps, include more variables than charts, often conveying a relational component. A Venn diagram is a good example of a graph. Its purpose is not so much to convey the relationship between two variables such as in a chart, but rather to convey the relationships between a variety of variables. A traditional map provides the spatial relationships between various points on the map.

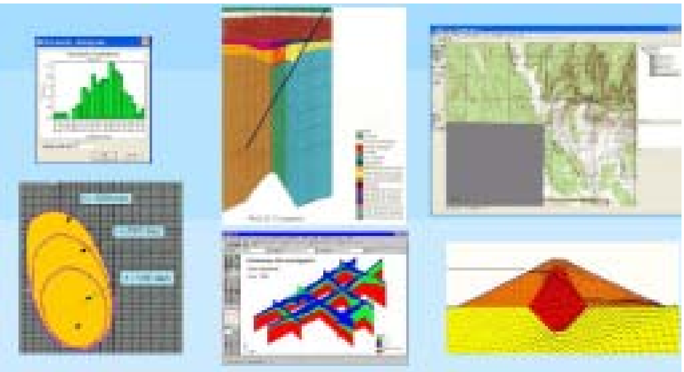


Figure 2. Examples generated using various visualization tools.

More recently, due to the rapid increase of power at all levels of computing, multimedia visualizations have become a practical reality. Multimedia includes movies, sound and other less traditional forms of visualization. One of the most interesting areas of visualization orbits around the power of using "traditional" data in new ways of display and analysis.

A rather recent view moves all visualization to the mathematically based idea of *mapping*. Mapping defines the algorithmic relationship of moving one or more data points onto a new surface. Generally, the mapping algorithm must maintain the data *topology* – the relationship of one "point" to all other points. A standard example of topology is that a coffee cup and a doughnut have the same topology.

While the number of database management systems and to some extent the number of decision support tools are limited, the number of data visualization tools is quite substantial. Although data visualization constitutes only a part of the entire analytical chain, it is generally not possible to entirely separate the visualization aspects of a data analysis from the analytical aspects. Rather than provide a subjective subset of tools, this overview will provide a structure to assist the reader with the selection process.

Considerations for Selecting Data Visualization Software

Although data visualization (DV) could easily tax a low end computer, generally speaking most business level hardware is capable of supporting a rich variety of DV software. This overview does not cover hardware requirements but as a generalization, hardware costs are not prohibitive. It is also true that there are hundreds if not thousands of visualization programs to serve many different users.

Some general considerations look at the process of bringing data into the system, filtering that data and then producing a visualization as an output product. Another set of considerations concerns the operation and maintenance (O&M) of the system over time. A third set of technical specifications is also provided.

General Considerations

- What types of data will the software import? Having more options for file import types will reduce the pre processing needed for the data package to be readily used.
- In conjunction with the first bullet, what types of GIS capabilities does the software have? Depending on the needs of the users, does the package meet the GIS needs of the user?
- Is the system user friendly for the intended user(s)? Depending on the expertise of the users in an organization, the package should be able to cater to their needs.
- How fast can the software render an image? Rendering a final image can be time consuming even on relatively powerful machines.

Operation and Maintenance Considerations

- Does the software operate within the current environment? Is the software compatible with the Operating System and various data sources? Do we need to invest in a new software or OS?
- Are the manuals useful? How well the documents are explained for a user to be able to make use of it?

- Is there training offered/required? If the training is needed and is not included in the package, this may end up being an expensive part of the package.
- Is the software support available and reliable?

Technical Considerations

- Does the software provide 2D, 3D, and 4D capabilities? See (http://research.esd.ornl.gov/CRERP/VISUALIZ/INDEX.HTM) for an example of 4D visualization. Not every application requires all types of visualizations.
- What is the largest image size possible? Some images can be very data dense.
- Will the data be available to special needs users?

Attachment

The attached table shows a comparison of several data management packages. These were selected from internet searches and probably do not include all the programs out there. Main criteria used in selecting these programs are: The program must have specified and defined fields, it should be easily available, transferable and ready to be used by users who can download it. The table shows, for each program, the operating system on which the program runs, number of tables that are in each program, the number of defined and the number of undefined fields. For each program, a general comment is provided that gives a brief description of the program, its capabilities and its limitations. Of course, these are only programs for data management and similar tables can be generated for data visualization and data analysis packages. As mentioned earlier, there are many more packages that can be compared for analysis and visualization. The RPO team is planning to maintain the list on their team website. However due to the page limitations within the overview those others are not included here. A good resource that compliments this document is available at EPA's website (EPA, 2004) on decision support tools, where a matrix for the data management is provided.

Summary & Conclusions

In this document, the ITRC-RPO team has made an attempt to put together introductory information on the DMAV tools. This is intended as a first source on these topics and is aimed at educating the state, federal, and other private agency regulatory and project management professionals who are interested in RPO and DMAV tools in particular. The importance of these tools is becoming more obvious in determining and implementing more efficient cleanup decisions at contaminated sites.

Contact information

Richard Hammond
United States Environmental Protection Agency
(404) 562-8535
Richard.Hammond@epa.gov

Sriram Madabhushi
South Carolina Department of Health and Environmental Control
(803) 896-4085
madabhs@dhec.sc.gov

References

Environmental Protection Agency. 2004. Matrix is available on the internet at: http://www.frtr.gov/decisionsupport/FunctionalGroups/Data_Management.htm.

Environmneetal Protectiona Agency. 2005. *Staged Electronic Data Deliverable*. Available on the internet at http://www.epa.gov/superfund/programs/clp/sedd.htm.

WERD - Western European Road Directors. 2003. "Data Management for Road Administrations: A Best Practice Guide." Page 3.

Program Name	Operating System	Number of Tables	Number of Defined Fields	Number of Undefined Fields	Comments
ChemPoint www.pointstar.com	MS Windows 95/98/2000/NT	0	161	0	Entry level environmental data management system with easy data import/export and reporting.
ChemPoint Pro www.pointstar.com	MS Access or Microsoft SQL Server	6	216	30	Full featured open-architecture client-server environmental data management system, with extensive data import/export, graphics, contouring, and reporting, with support for third party product integration.
EnABL www.enabl.com	Oracle with a web browser access	493	2327	1933	The enABL Data Management System (EDMS) is a state-of-the-art Internet-accessible Oracle-based system, providing multiple parties access to high-integrity, fully validated environmental data. The EDMS captures and maintains analytical laboratory data with information related to sites, locations, soil borings, lithology, well installation and monitoring, field sampling and field testing data, chain-of-custody data, data quality objectives, and regulatory requirements.
EQuIS 5 www.earthsoft.com	MS Access, Windows 95/98/2000/NT, MS SQL, Oracle	136	1386	Variable depending on system	Simple enough for one facility yet scalable to thousands of facilities, EQuIS 5 combines the power of a desktop application (EQuIS Professional) with the automation and ease of use of a web based application (EQuIS Enterprise) both drawing from the same database. EQuIS 5 is developed in an MS .NET framework and available in MSDE, SQL Server, or Oracle versions.

Program Name	Operating System	Number of Tables	Number of Defined Fields	Number of Undefined Fields	Comments
EQWin www.gemteck.com	MS Access, MS SQL Server	12	101	0	EQWin Data Manager is a desktop and/or enterprise level database application that handles many different environmental data types, including surface/ground water, biological, sediment, soil, and air quality data. The system is designed to ensure environmental sampling data validity, as well as to store, analyze and report on an industry's impact on the environment against regulatory and/or permit standards.
Enviro Data www.geotech.com	MS access	126	1931	0	Clients report that Enviro Data is the most cost-effective and easiest to use system for managing environmental lab and field data. Complimentary tools extend Enviro Data to provide planning, tracking, and GIS display capabilities.
HoleBASE III www.keygeotech.co.uk	Windows (98 to XP Pro)	80	1123	254 max/ table	HoleBASE and its companion applications (collectively called 'The Geotechnical Office') deal with the whole range of site investigation data management issues, from initial data acquisition, through processing and production of reports and analysis graphics, to powerful ground modeling and 3d visualizations. In addition, HoleBASE makes data available to other applications, thereby ensuring that users can always use the most appropriate tool to achieve their data presentation goals.

Program Name	Operating System	Number of Tables	Number of Defined Fields	Number of Undefined Fields	Comments
Hydrogeo Analyst www.waterloohydroge-ologic.com	MS SQL server	43	352	Un-limited	HydroGeo Analyst offers the ultimate level of database flexibility and accomplishes this by giving user's the ability to add tables or fields to the database at any stage during the project. HydroGeo Analyst's core is built on Microsoft SQL technology, and is coupled with an extremely user-friendly graphical user interface (GUI) that instantly adapts to any changes to the data-base model/structure.
Locus Focus Environmental Information Management www.locustec.com	MS SQL Server w/ web access	274	3091	23	EIM is an entirely web-based enterprise system for uploading, managing, and viewing sampling, analytical and other environmental information pertaining to site investigations, cleanups, and ongoing 0&M. A robust electronic data validation module, a flexible EDD loader, Quick Views for data mining and rapid reporting, and eGIS-SVG, a powerful module for viewing data on maps using Adobe's free Scalable Vector Graphics Viewer, are some of the many features available in EIM.
Terrabase www.bossintl.com	Windows 98, NT, 2000, XP, MS Access, MS SQL Server	110	2000	0	TerraBase is an intuitive, SQL-compliant relational database application designed for environmental pro-fessionals and managers who need to assess and manage chemical, geological, and spatial data at one or more sites. TerraBase is equipped with powerful reporting capabilities allowing the user to print spatial and tabular results in pre-formatted reports as well as output data queries automatically to other applications including spatial analysis tools.

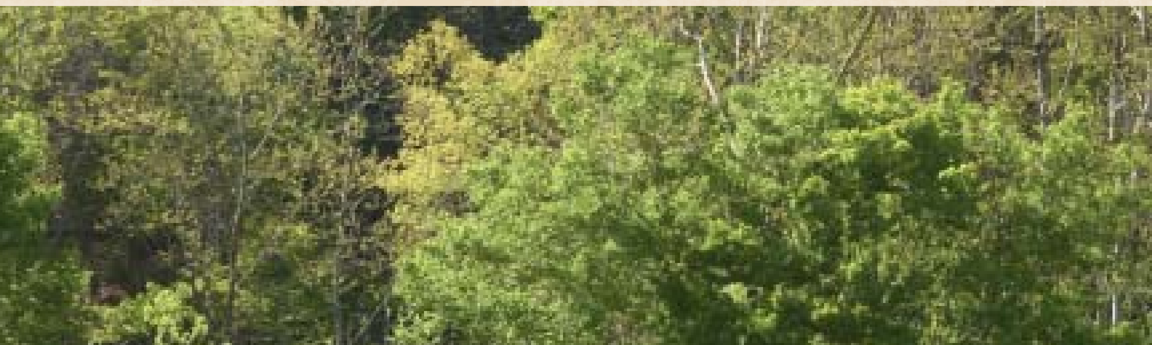
Program Name	Operating System	Number of Tables	Number of Defined Fields	Number of Undefined Fields	Comments
Visual Site Manager www.geoanalysis.com	Windows 98, NT, 2000, XP, MS Access	60	400	40	Visual Site Manager TM (VSM) is a DMS that provides a user-friendly, map-based interface to chemical, hydrologic, and geologic data from environmental and mining projects. VSM is designed to integrate existing CAD maps, photos, and geologic logs with a open MS Access database.

Criteria for selecting programs:
Had to have specified and defined fields.
The program has to be easily transferable.

Disclaimer: The programs were found as a result of internet searches.

The above programs were selected as data management software packages only. Visualization, analysis and other capabilities were not considered as criteria for selecting these programs.

www.itrcweb.org



RPO-5

Contacts:

RPO Team Leaders

Sriram Madabhushi, P.G.
SCDHEC
2600 Bull Street
Columbia, SC 29201
803-896-4085
madabhs@dhec.sc.gov

Tom O'Neill
NJ Department of Environmental Protection
P.O. Box 413
401 East State Street, Sixth Floor
Trenton, NJ 08625-0413
609-292-2150
tom.o'neill@dep.state.nj.us