

Managing Environmental Data

Laurence Davidson
EarthFx Inc.

Lisa Keller
Environment Canada

SUMMARY

Advances in information technology software have enabled considerable progress in environmental data management techniques. Desktop relational databases are rapidly replacing spreadsheets for data storage, and a variety of data management systems are emerging to streamline the process of integrating and mapping data, with some including basic logic sequences providing decision support functionality. Simultaneously, we are also collecting more information as data logger technology now enables remote satellite links for virtually continuous data streams of water level, water quality data and the like.

As a profession, we should consider becoming more efficient in the ways we manage our data. The rewards are considerable.

- More effective use of our time as environmental scientists
- Preserve the value of the data for future sale and cost recovery
- Optimize our use of application-specific environmental software appearing on the market

Following is an explanation of the importance of data management, and descriptions and examples specific to the environmental sciences. It is intended that this collective wisdom is sufficient to serve as a guide to those undertaking an environmental database for the first time, yet also offer insight for more advanced users.

INTRODUCTION

Once contamination is suspected at a site, we embark on a two-step process to manage the site, and ideally return it to a reasonable land use. The first step is to collect data from the site to characterize the geologic/hydrogeologic conditions, and quantify the extent and composition of the contaminants. Based on these findings, the second step involves making decisions on what action is necessary; do we remediate, and how? – or is monitor/natural attenuation the most cost effective solution with the current technology pool, or is more study necessary?.

To make these decisions, we often make use of a variety of tools, including Decision Support systems to provide a consistent and reproducible evaluation and sensitivity analysis of remedial options. This is particularly important in light of the range of parameters necessary for such decisions, including economic, social, technical and environmental factors, and all their associated data. The annual variation in water table is an example of a common technical input variable that

can be easily queried from a database. Data management therefore becomes important component of a defensible DS system. This is further emphasized by considering the types and sheer volume of data required. Using a DNAPL in fractured rock site in Canada as an example, the following lists some of the data collected (the site has benefited from considerable research work, and thus offers an unusually large suite of data and may not be representative of more typical projects).

Data Type	Purpose	Volume
Boreholes	Geologic & hydrogeologic structure / composition	<ul style="list-style-type: none"> • 170 boreholes • 500 measurements of stratigraphic contacts • 20,000 fracture occurrences and other minor geologic descriptions
Monitors	Ground water flow and quality measurements	<ul style="list-style-type: none"> • 470 monitors • 600,000 water level measurements • 160,000 water quality measurements
Core Photographs	Enhanced geologic interpretation and fracture mapping	700 core photos
Packer Data	Packer tests completed in about 10 percent of wells to further define vertical permeability profile..	Between 1000 and 2000 individual permeability readings
Survey Data	Consistent reference point elevations	400 measurements

This tabular data creates a database file likely exceeding 200 mega bytes, with a possible additional 800 Mbytes of core photographs, site plans, air photos and borehole geophysical logs. If printed, the tabular data alone would occupy in excess of 15,000 pages. Understandably, a database system designed for rapid searching of such data becomes an important tool for satisfying the data needs of Decision Support Systems and other such users of the data.

OUR CURRENT POSITION

*Are you getting everything out of your earth science data? It sure costs a lot to collect.
 Are you spending more time organizing data for your report than for your analysis?
 Are you able to merge the various data measurement types into an integrated interpretation?*

These questions highlight the common shortcomings of traditional environmental data management, typically based on systems where data was digitized for display not storage, and where paper reports were the final project deliverable and often the only repository of the data.

This is changing. The environmental earth science community is undergoing a significant revolution in how we manage our data, stepping from the 'report-centric' to the 'data-centric' world where the final deliverable of a project is a CD containing the project database.

This transition is occurring because of independent forces coming to bear on how we work. First is the global downsizing and the need to do more with less. Tools that promote our efficiency are

obviously welcome under these conditions. Second is the rate as which we now collect data. The use of data loggers, quantitative broad analytical spectrum of data from the labs and GPS units have dramatically increased the volume of data available for analysis in any given project. Third, the rapid growth in processor speed and software have now brought mainframe database computation power to the desktop computer. And finally, accreditation programs, such as ISO 14000, promote environmental awareness and auditing within firms, a driving force for software development.

To capitalize on, and arguably survive, these trends, environmental scientists and engineers are adopting more rigorous data management methods. As learned from the petroleum geoscience sector, this encourages two important outcomes. First, it will enhance our ability as earth scientists to analyze and understand the physical and chemical systems bearing on our projects as we are more efficient and storing, retrieving and analyzing data. Second, it will promote the value of the data itself. With proper storage, the data has value to others, offering means of offsetting the initial collection costs.

However, such advances come with new challenges. There is a need to incorporate basic database design constraints to ensure flexibility in our systems. There is a need to give special consideration to the scale at which we want the data to perform. There is a need for data models that establish standards for how data is stored, thus facilitating the sharing and sale of data. yet be implemented within a framework that recognizes that these models will evolve. And finally, there is a need to develop an understanding of how to judge, select and use the numerous new earth science software products reaching the market ever day, to ensure work processes are enhanced, not burdened by these tools.

As a starting point, we offer a snapshot of the data management characteristics of today's environmental data management methods.

1. In the absence of specific and effective environmental software, environmental scientists are famous for their ability to mix and match commercial software based on the data at hand and project requirements. Examples include CAD, GIS, SURFER, word processors, spreadsheets etc. The consequence is that project data is divided (and partially duplicated) across several software file formats, and several computers / servers.
2. The problem is further compounded by the use of software that uses proprietary internal data storage, often with limited import/export functions. Such 'data holes' conflict with most data management concepts, and should be avoided.
3. It is not uncommon to see both consultants and manufacturers developing internal or 'home grown' data management/analysis systems. However, given the growing recognition that software is a service and not a product, such firms are often overwhelmed by the demands of training, supporting and improving their software. This trend was experienced in the late eighties / early nineties in parts of the petroleum sector where consulting firms instituted software development departments. Most such departments have now been eliminated as the firms focus on their core business.

4. Enterprise scale database consultants have been retained to design and build proprietary environmental data management systems for large corporations or municipalities. In our experience, this is not necessarily a full proof approach, as demonstrated by a Canadian municipality that was forced to mothball elements of its new system when it was discovered that the data model was unable to handle hydrogeological queries. The database designers had only a limited understanding of hydrogeology, and did not appreciate the business rules of their client.

These examples serve to underline several important aspect of environmental data management.

- *Data storage and data analysis are separate tasks, and should be handled by separate software.*
- *A central, spatial, open, earth science– relational database is a fundamental foundation of any system that manages data.*
- *The selection of application specific environmental software is growing rapidly. It is a buyers market. Use this to your advantage by having your data readily available in a proper database.*

SYSTEM CLASSIFICATION

In order to discuss data management aspects of environmental software, there is a need to establish a loose classification framework to establish the relative roles of software packages. A DMS classification grid is provided in the following figure.

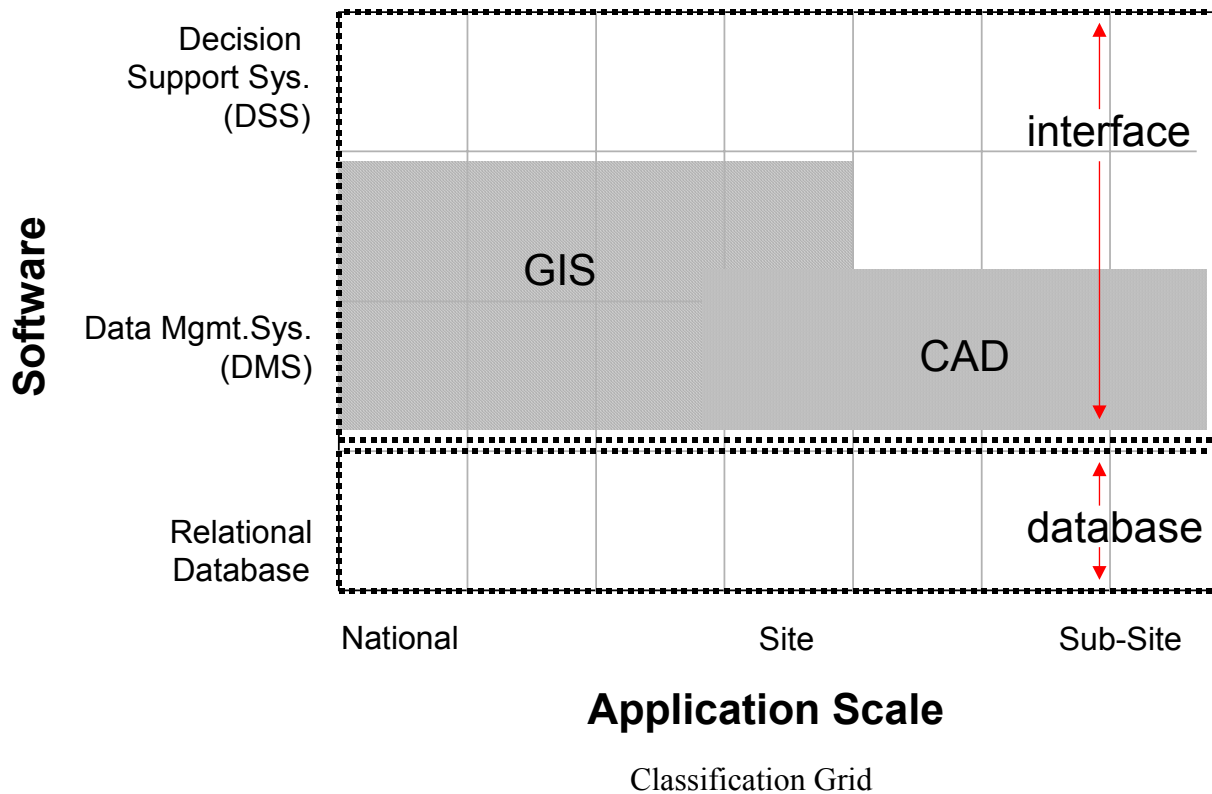
The horizontal axis represents the scale of the application, in this case ranging from sub-site to national. There are substantial differences in data needs and data structure between systems that are designed to optimize bio-remediation system, and a national-scale application, used to track and manage hundreds if not thousands of individual sites.

The vertical axis represents the degree of built in logic offered by the system, ranging from none in a database, to a Decision Support System. As further definition, the intermediate DMS would include functions for managing and validating data and seeking and displaying data. GIS offers much of the later, but little of the former. DSS is used to describe systems that have the ability to show ‘derivative data’, either by krigging surfaces, assembling data objects for advanced perspectives, or applying established business rules and user defined criteria to support decisions.

Several issues to consider on the grid include:

1. The boundary between database and ‘interface’ (used to represent the spectrum of DMS, DSS, analysis and display software) is clearly defined. Software products should not cross the boundary, as this will likely entail some form of proprietary data storage on behalf of the interface.
2. Other axis could be added to the grid to illustrate the mapping capabilities of the interface, and web friendliness, as two examples of critical features of such systems. Users are strongly encouraged to consider these options during any evaluation process.

- To serve as examples, the realms of typical CAD and GIS packages are shown. CAD largely as a site to sub-site application, and GIS as a site to multi-site application. Note that although some GIS offer internal data storage, they also offer ODBC connections to external databases.



DATABASE DESIGN

The environmental database should be separate and independent from other project software, be based on commercially available software and use a data model that recognizes the relationships between the data being collected (e.g., water levels from a monitor and geological descriptions from the borehole).

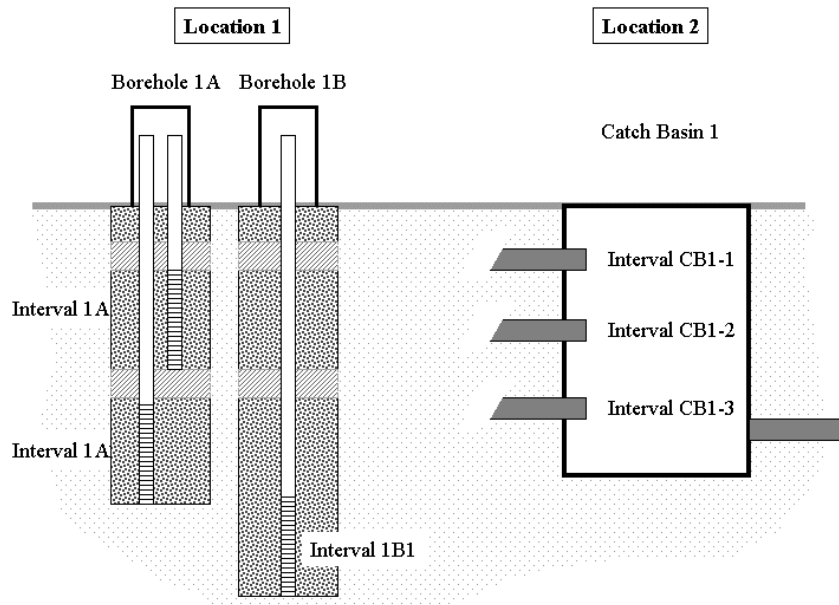
Step 1: Naming Conventions

When first designing the database, it is important to name and classify the various tangible objects from which we collect data (such as boreholes, groundwater monitors, soil samples, sample jars for lab analysis, rain gauges, etc). The naming should be sufficiently generic to ensure an element of scalability to the database. We have found the Location / Borehole / Interval system to work well. An example is provided below.

TIP

Naming Convention when designing database

- develop a hierarchical list of the objects you foresee collecting data from
- assign generic names to these objects
- carry these names through the database, both in table names and field names



Example Naming Convention

- Location:** Locations are the primary source of data and can be thought of as a spot on the map where we collect data from. The Location table contains the x, y and z coordinates and the location name.
- Borehole** These are secondary tables, and contain specific information about each borehole (and catch basins and other tangible objects that give rise to data). As an example, the Borehole table would contain the Borehole Name, the name of the location at which it is found, depth etc. Boreholes are handled as three dimensional objects, with coordinates for the top and bottom, dip and strike. The Borehole table would be linked to tables containing details of the sand-pack, seals, casing etc.
- Interval** Intervals are also three dimensional objects, and can be thought of as the groundwater monitoring screen or stream flow gauge or the like, basically a point in space to which we can return to collect temporal data.

Step 2: Data Model

As earth scientists, the vast majority of our data comes from ‘the vertical dimension’ in the form of test pits, drill holes, test wells etc, rather than from a ‘surface’ or map. It is important to build this dimension into the data model, at a low level by ensuring that each data point has a reference in

space (x,y,z) and time (t), and at a high level in terms of the relationships between the database tables. This section will focus primarily on the relationships between tables, with the content of the tables discussed in the following section.

Developing an appropriate data model is critical to the long-term success of the database as it establishes limits on the following measures of effectiveness of the database. The data model:

- Should not limit how data can be queried
- Is scalable and accommodates growth of the database in the form of new data tables
- Can be easily linked to other similar databases

A good data model will not constrain how data can be queried (for example, a poor model may not recognize monitors as separate objects from boreholes, and thus disable the ability to query water levels in a particular geological horizon), is scalable, meaning that new data type can be easily added (as an example, precipitation data from a rainfall gauge added to the site monitoring network can be easily accommodated) and finally, the database can be seamlessly linked to other similar databases, enabling for example, a comparison of TCE levels at various sites.

In North America, a current and growing hydrogeological database challenge is developing data models that accommodate similar data, but from different scales. A case in point is in Ontario Canada where work is underway to combine the provincial oil and gas well database (deep wells with some borehole geophysics and good bedrock definition) with the provincial water well database (over 600,000 private water supply wells providing good regional scale definition of water resources and overburden / shallow bedrock geology) with geotechnical / hydrogeological test wells largely installed by the consulting community, offering high quality geological and temporal (water levels and water quality) data. The scale at which data is collected from these three well types is considerably different, invoking challenges in developing a common data model, but very much in demand in light of the power of such an integrated data source. The issue of scale is discussed in further detail later in this document.

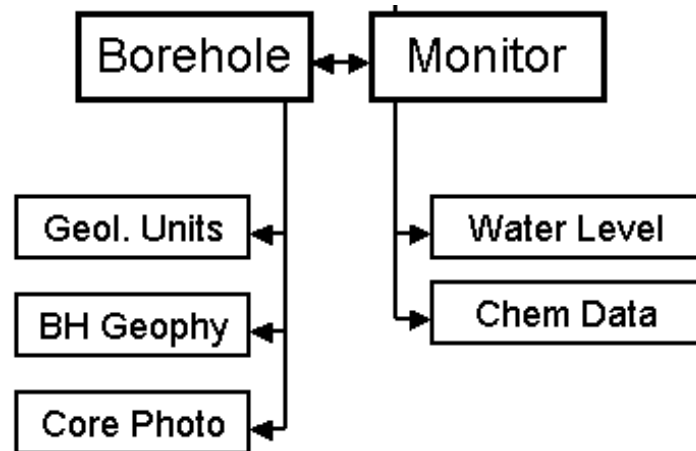
It is noteworthy that the petroleum sector has developed extensive data models to serve as templates for their members in an effort to promote a 'data standard' across the industry. Examples of the petroleum data models can be found at www.posc.org or www.ppdm.org, and a visit is encouraged for those developing environmental / hydrogeological databases.

Example Data Models

Two data models are presented here for consideration. The first is simplistic, but conveys the basic design elements of an environmental / hydrogeological database, the second is an expanded version of the first, and displays most of the key tables and relationships of a generic data model currently in use with the authors. Functional examples in MSAccess of this model will be available to download from www.earthfx.com.

Simple Data Model

The adjacent figure illustrates the basic data model, with individual tables shown as boxes, and the lines representing the links between the tables. Key elements are the borehole – monitor relationship to which all data is connected. Intuitively this model applies well to traditional environmental monitoring wells, using screens (or monitors) to measure groundwater levels and quality. Drilling details such as depth, date contractor dip etc are stored in the BOREHOLE table,



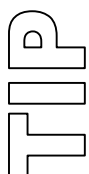
details of the monitoring screens (depth, slot size, material etc) are stored in the MONITOR table. Information from the borehole, such as geological descriptions, borehole geophysics core photograph details etc) are stored in separate tables, linked to the borehole table. This structure begins to fail should we consider adding climatic data, as the details of meteorological station bear limited similarity to a borehole, and thus does not ‘fit’ into the BOREHOLE table.

Current Model

To accommodate the range of data collected during environmental investigations, a more versatile and generic data model is necessary. The authors suggest the model shown in the following figure, now currently in use in several large Canadian environmental databases. The model has adopted a generic nomenclature (as discussed earlier) for monitoring stations, using Locations to refer to the map location where data is collected. It could be a single well, a cluster of nested wells, a catch basin or climate station. At a given Location, there are specific instruments or intervals from which we collect data. Examples include a thermometer at the climate station, a screened monitoring interval at a well or a split spoon sample from a test boring. These data sources are called Intervals. The LOCATION and INTERVAL tables are loosely referred to as Level 1 tables as they form the fundamental source of all data in the database.

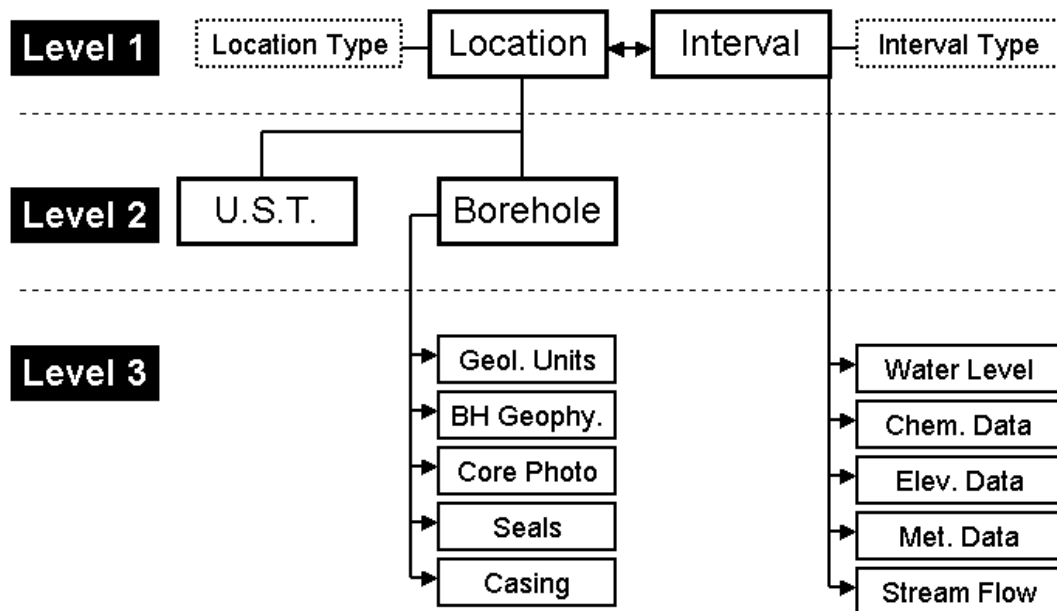
To accommodate specific detail about the drill hole, climate stations, Level 2 tables are necessary. In complex site investigations, Level 2 table such as Borehole, UST (underground storage tanks), Climate (weather station details), Stream (river gauging station details) would be expected. As with the Simple Model, Level 3 tables containing the actual field measurements are added.

This model has shown itself to accommodate a wide variety of data types collected, and from a range of scales.



Designing The Data Model

- With pen and paper, draw a road map of the planned database showing the tables and their links, considering all possible data types
- Write down what data will be stored in which tables.



Example Environmental Data Model

Step 3: Table Design

Tables are where the data is stored, and resemble spreadsheets, with columns and rows. A database is made up of one or many tables, each storing ‘one type of thing’ within the database (i.e., water levels, or chemistry data).

When designing databases, decisions are made regarding how many tables, what they will contain, how are they linked. The answer to these questions lies in something called normalization, the process of simplifying the design of the database to achieve optimum structure and hence performance.

Normalization theory includes the idea of normal forms to assist in achieving optimal structure. The normal forms are a linear progression of rules applied to the tables, with each higher rule achieving a more effective design. Five levels of normal forms are recognized, in practice, we try and adhere to the first three, and sometimes settle for two. A brief description of the normal forms follows.

Before first normal form – relations

To apply the normal forms, tables must meet the following criteria

- They describe one subject or entity
- They have no duplicate rows
- Columns and rows are unordered

First Normal Form (1NF)

A common example is people's names. Following 1NF, the first and last names should be stored in separate columns, where traditional spreadsheet thinking might place them together.

Second Normal Form (2NF)

Tables should describe only one 'thing'. For example, a table of borehole information should not contain the address of the drilling company as it is the same for all boreholes, and thus should be stored in a separate table.

Third Normal Form (3NF)

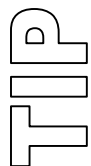
An example is calculated columns. A table of monitor details may include the depth to the top of the monitoring interval, but should not include the corresponding elevation, as this can be simply calculated each time needed. (it is noteworthy that in practice, especially in large databases, it is more efficient to include calculated columns that are referenced on a regular basis as the computational time to perform the calculations can become extensive).

Relationships

A fundamental component of relational databases is referential integrity. Referential integrity is applied to individual table to table relationships in the databases. When activated, referential integrity prevents 'orphaned data', or as an example, it prohibits a sample from existing in a Sample table if the monitor from which the sample was collected does not exist in a Monitor table. This is a very powerful feature, and in itself ensures a high level of quality assurance in the data.

How does this theory translate into an environmentally friendly database? First, the following figure illustrates a sufficiently normalized table of chemistry data. The key element of this design is that each record, or row, contains only one chemistry result, identified with a sampling location (IntName), sampling date (CDDate), parameter, value, units, method detection limit (MDL) and something called ChemModifier, a text string used to qualify the result. We have also added the identification number of the sampling location (IntID) and the sample matrix (set as 2 for soil). Naturally additional fields such as analysis date and the like can be added to the core fields shown. The same logic applies to water level tables, climate tables etc.

From an implementation perspective, it is becoming increasingly popular to define required data standards in the contracts or Terms of Reference to laboratories and other contractors who will be submitting large data sets. The standards define the structure, field types and any required headers, and ensure that the data can be seamlessly appended to the database. The agency responsible for the database should aim to minimize any restructuring or retyping of third party data as this entails a higher degree of responsibility for the data, and consequently a manual checking process.



Designing Normalized Tables

- It is important that the temporal data tables (water level, climate, chemistry) be properly normalized
- Database flexibility and functionality will otherwise be lost

IntName	IntID	Matrix	CDDate	Parameter	ChemModifier	Value	Units	MDL
BH14-SA1	1053	2	9/20/91	pH		85.10001		
BH14-SA1	1053	2	9/20/91	Selenium	<	1	ug/g	0.1
BH14-SA1	1053	2	9/20/91	Silver		7	ug/g	1
BH14-SA1	1053	2	9/20/91	Titanium		600	ug/g	
BH14-SA1	1053	2	9/20/91	Vanadium		200	ug/g	1
BH14-SA1	1053	2	9/20/91	Zinc		630	ug/g	0.5
BH14-SA2	1054	2	9/20/91	Antimony	<	8	ug/g	0.8
BH14-SA2	1054	2	9/20/91	Arsenic		57	ug/g	0.5
BH14-SA2	1054	2	9/20/91	Barium		650	ug/g	0.5
BH14-SA2	1054	2	9/20/91	Beryllium		5	ug/g	0.5
BH14-SA2	1054	2	9/20/91	Boron	<	100	ug/g	10

Example Normalized Table Structure

When assigning data to each table, remember that a good database will have each value stored only once (much easier for corrections) and made available throughout the database using relationships. As example, the elevation of the top of the monitor is stored in only one table, but may be accessed, via the relationships, by data in many other tables, converting water depths to elevation, for example.

Step 4: Special Fields

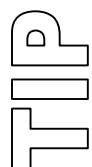
Calculated Fields

Database theory dictates that values are stored in their raw form, and any calculations are performed at run time using queries to feed printed reports or on-screen form. This is where issues of performance and convenience supercede the theory. We have found that OUOM fields (Original Unit of Measure) indispensable for maintaining data integrity. Each field containing a measured value, such as water level or chemistry data parameter, value and unit, should occur twice in the table. The first occurrence holds the measurement as provided to the database, the Original Unit of Measure. This field name can be suffixed with OUOM for identification. The second occurrence will hold the *standardized* equivalent of the OUOM value. Although this amounts to duplication, the benefits are significant.

Consider chemistry data. Two laboratories send you weekly data files that you merge into the database. The first laboratory reports soil chemistry as ppm, and uses the trichloroethylene naming convention. The second lab reports in ug/g, and has adopted the trichloroethene naming convention. We know that these are the same compounds and the same units, but the database does not, nor does any software being used to analyze or display the data. Significant errors therefore occur in any query seeking trichloroethene. By using OUOM fields, the original lab data is preserved intact for future reference and quality control, and queries update the standardized equivalent on a periodic basis. OUOM fields are also useful for auditing and ISO purposes.

IntName	CDDate	ParameterOUOM	ValueOUOM	ChemUnitsOUOM	MDLouom	Parameter	Value	Units	MDL
BH12-SA1	9/20/91	Antimony	8.1 mg/kg		0.8	Antimony	8.1 ug/g		0
BH12-SA1	9/20/91	Arsenic	280 mg/kg		0.5	Arsenic	280 ug/g		0
BH12-SA1	9/20/91	Barium	440 mg/kg		0.5	Barium	440 ug/g		0
BH12-SA1	9/20/91	Beryllium	2 mg/kg		0.5	Beryllium	2 ug/g		0
BH12-SA1	9/20/91	Boron	53000 mg/kg		10	Boron	53000 ug/g		1
BH12-SA1	9/20/91	Boron (available)	19000 mg/kg			Boron (available)	19000 ug/g		
BH12-SA1	9/20/91	Cadmium	4 mg/kg		1	Cadmium	4 ug/g		
BH12-SA1	9/20/91	Chromium	330 mg/kg		1	Chromium	330 ug/g		
BH12-SA1	9/20/91	Chromium (VI)	mg/kg			Chromium (VI)	ug/g		
BH12-SA1	9/20/91	Cobalt	45 mg/kg		4.5	Cobalt	45 ug/g		4
BH12-SA1	9/20/91	Copper	210 mg/kg		1	Copper	210 ug/g		
BH12-SA1	9/20/91	Lead	810 mg/kg		10	Lead	810 ug/g		1

Example of Original Unit of Measure Fields



Keeping Original Data

- Duplicate fields should be created for all measurement, one for the value as recorded, the other for a transposed or standardized equivalent
- This is important for quality control, data integrity and auditing

Unit Fields

Closely associated with the OUOM fields are units fields. As shown in the above figure, the units of every measurement are stored in their respective fields, for both OUOM and subsequent calculated fields. Although this may appear at first glance as an un-necessary use of space, and rigorous database theory would suggest that all measurements be taken with the same set of units, the authors suggest tracking all units. In practice, measurements are not collected with consistent units (laboratories commonly mix ppm, mg/L, ppb in a single report) and serious errors in analysis occur if units are overlooked.

Tracking Site Identification

For environmental/hydrogeological databases, tracking the site where each Location is located adds an important scaling factor to the database. In our experience, many small but good databases grow to accommodate multiple sites. Build this option in at the start by tracking the site for each Location, and all water level records, chemistry data records and other such interval data.

Relative Quality Fields

Not all data is collected equally. The elevation of the top of a well may be surveyed against a known benchmark, and thus have an error margin of less than 0.01 m, or it may be estimated from a topographic map with a margin of 5 meters. Similarly, laboratory data may be derived from accredited laboratories, field laboratories, or field assay kits, be diluted, or reported at or near method detection limits, all of which control the amount of confidence we can associate with the reported value.

Relative quality fields are therefore commonly used for data where the error margin can vary significantly, allowing users to selectively query based on data quality, if necessary, or build the error margin into any subsequent geo-statistical analysis. Coordinates and elevation are common

examples, as is laboratory data. The Relative Quality field is often linked to a ‘Look-Up’ table, listing a range of pre-defined error categories to choose from. An example is shown below.

Error Code	Description
1	Surveyed in field from known Bench Mark. Instrument level, accurate to 1 ft.
2	Surveyed in field from known Bench Mark. Instrument level, accurate to 5 ft.
3	Surveyed in field from known Bench Mark. Instrument level, accurate to 10 ft.
4	Elevation read from topographic map, contour interval - 10 ft.
5	Elevation read from topographic map, contour interval - 25 ft.
6	Elevation read from topographic map, contour interval - 50 ft.
7	Elevation read from topographic map, contour interval - 100 ft.
8	Elevation read from topographic map, contours crowded, i.e. location point of well touches on 2 or more contour lines.
9	Elevation accuracy unknown or unreliable. Leave elevation blank.

Example Look-Up table for Relative Quality Fields

Step 5: Issues of Scale

Databases inherently have scale. Scale is manifested through the types of data fields included in the database (such as soil porosity or hydraulic conductivity, measurements that apply only on a small scale and cannot be extrapolated over a distance) and the fields used to position the data on map. Examples include fields such as the Site field, plus the use of latitudes and longitudes as well as UTM, and / or the use of a UTM zone field.

Inconsistent database scales complicate merging and linking of separate databases. Imagine trying to merge a database of oil wells in the North Sea with a database of monitoring wells installed around a service station. There would be few, if any, common elements. This may be an extreme example, but when designing a data model, think of this example, and include fields to accommodate the scale spectrum.

OUTCOMES

Proper data management is a means to an end. It is a tool that allows us to work smarter, producing a better product in less time. As environmental scientists, data management systems give us immediate access to our data (offering greater opportunity for more advanced analysis), and as a result, instill greater confidence in our analysis and decisions. The following aspects of data management contribute to this outcome.

Quick Results

Structured Query Language or SQL is a simple computer language user to query data. It is the engine of the database, allowing hundreds of thousands of data records to be searched in a fraction of a second. Examples of typical SQL requests (or queries) include:

- Average water level in the overburden in June 2000

- Maximum and minimum water levels at the site
- All wells were water quality exceeds regulatory criteria

Desktop databases provide intuitive tools allowing users to rapidly develop complex queries and thus dramatically accelerate the process of searching for anomalies and trends in the data. In addition, using SQL, explicit datasets can be queried and exported to third party software for further analysis, such as interpolation packages and CAD/GIS.

Improved Data Quality

Databases and data management systems offer improved quality control over traditional spreadsheet data storage.

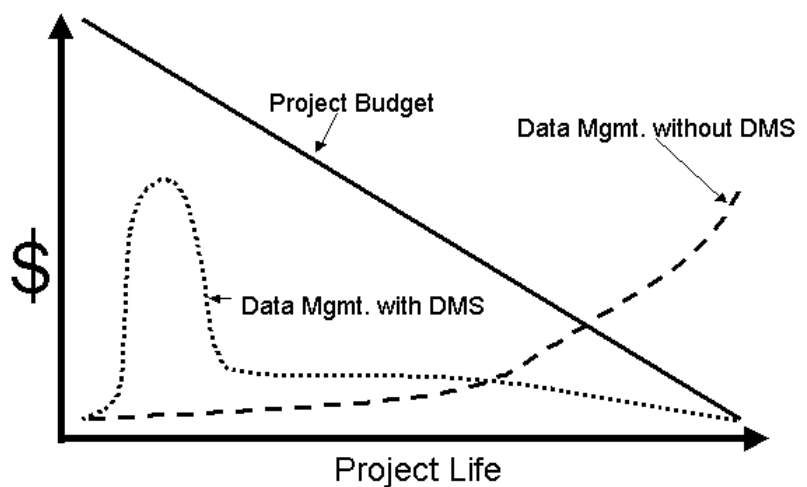
One Copy Storage All data is stored in a single database file that serves all project staff. Data is not duplicated across multiple computers, eliminating confusion with corrections and updates. In addition, the database structure ensures that data are stored only once, simplifying any required updates or corrections. The top of well reference elevation is a good example.

Business Rules Basic business rules can be built into the database design to minimize data entry errors. Examples may include preventing the entry of a water level for a well that does not exist.

Authority Databases are often managed by an individual or a small team who coach novice users, prepare standard queries and reports, look after back up issues, build in business rules and other tasks that protect the value of the data. They also carry out standard database administration tasks, such as checking that units are consistent.

Reduced Cost

Implementing a data management system is an up-front cost to a project. The expense occurs as the start of the project, when the demands on the system are minimal, and when sufficient budget is available to test the system and train staff. The operating cost of the system is small, producing large returns on the investment in the final phases of the project when data demands are at their peak, and the budget largely expended. This pattern is shown schematically with traditional data management costs in the accompanying figure.



Integrated Analysis

With all data stored together, we are able to integrate results from various data collection streams (water levels, geology, water/soil quality) and perform more complete and effective analysis. This compares to searching through countless spreadsheets, followed by a cut and paste operation and re-formatting to assemble the required data. Not only is this a time consuming exercise, to be repeated each time the original spreadsheets are updated, but it offers potential to introduce error into the data.

CONCLUSIONS

1. The environmental sector often overlooks the value of the data collected, relegating data to paper appendices where it is effectively lost.
2. The widespread use of open, relational databases for environmental data management is recommended. Spreadsheets and 'proprietary data storage systems' of several analysis and mapping applications are to be avoided.
3. Databases promote the value of data. This occurs through their inherent validation process, 'one copy storage of values', structure and querying capabilities. Data stored in such a system has value to others, and can be sold to offset collection costs. The petroleum sector has clearly demonstrated this trend.
4. The design of the database is critical to its long-term viability. When designing a database, consider all the tangible objects that will yield data, and ensure they have a place in the database (examples include boreholes, monitors, sample jars, rain gauges etc). Consult with those who will use the data on a regular basis during the design. Carefully consider issues of scale, both in terms of the fields for the data being collected, and for fields used to position the data on a map. Also think about future links to other similar databases, such as national oil and gas well databases as a large-scale example.
5. The availability of application specific software for the environmental market is growing rapidly. Analysis systems, mapping packages, numerical models, criteria filtering, decision support and others are more common, more user-friendly, more applicable, and more economical. Market forces will compel use to use them. Relational databases will be the common link between the software packages, and underpin our analysis and understanding of environmental matters.