

DATA MANAGEMENT FOR LARGE VOLUME DATASETS

by

T.E. Rymer
J.J. Renton
A.H. Stiller

INTRODUCTION

Under contract from the W. Va. Department of Energy, the West Virginia Geological and Economic Survey is compiling an AMD database containing the geological, geochemical, hydrologic, reclamation and engineering information available in mine permit applications of all surface mines still under bond in the 18 counties of northern West Virginia. The database is to provide the basis for a research effort to better understand the parameters that control the rate and level of acid production on both a stratigraphic and geographic basis. It is hoped that such an understanding in conjunction with current research into acid prediction and ameliorization will provide adequate pre-mine engineering/chemical design parameters to eliminate the production of acid in future mining operations.

DATA COLLECTION

To date, the information from 344 surface mine sites has been acquired from D.O.E. files and has been computerized. These data fall into six basic data groups:

1. General mine information including location, seam(s) mined, area, mine type, date opened, overburden thickness etc.
2. Acid-base account data for all cores taken at the site
3. Soil information including type, surface slope etc.
4. Surface reclamation including, topical additives (fertilizer/lime), plants used etc.
5. Water sampling sites including location, type, distance from mine etc.
6. Water analyses

At the time of this writing, the data occupy more than 8 million bytes of computer storage and has consumed over 3 days in central processing unit (CPU) time during processing.

DATA MANAGEMENT

This vast amount of information demands that the processing, searching, statistical treatment and display of the data be accomplished by high speed computer. This in turn necessitated the design, testing and implementation of a database management system which would not only provide easy access to the data but would also provide the means by which the data

could be utilized to investigate and solve the problems of acid mine drainage formation/ameliorization. To this end, the COAL RECLAMATION INFORMATION SYSTEM (CRIS) software package was developed.

COAL RECLAMATION INFORMATION SYSTEM (CRIS)

The CRIS software package provides three basic operations:

1. file management which allows the user to search the database for information of specific interest
2. provide simple statistics for water or overburden data or to generate a working file of data which can be submitted to any other statistical software routine
3. the generation of a variety of graphical presentations including two or three dimensional plots and maps

At present, CRIS does not exist on a single computer but rather utilizes several different computing systems, transferring information along a network of computers with each individual computer storing descriptive information on database management software that is made available to CRIS. The data are compiled in the DIGITAL VAX/11 computer at the West Virginia Geological and Economic Survey. Using DEC DATATRIEVE, the raw data from the D.O.E. files are coded and transferred to the W.V.G.E.S computer which packs the information into binary arrays for efficient storage. The coded data are then stored on magnetic tape.

Three major files are generated from DATATRIEVE

1. MINES.DAT which includes permit no., geographical location, mining history, coal seam(s) mined, overburden thickness and any other information available in the permit pertaining to the mine site.
2. RCLMCHEM.DAT which includes all overburden analytical information and
3. BASELINE.DAT which includes all water quality data

These files are uploaded onto another DIGITAL VAX/11 computer at the Department of Geology, West Virginia University where most of the software programs are utilized.

Because the database is constantly being updated with information from new mine permits, software known as UPDATE was written which appends new data to the existing database. In addition, UPDATE utilizes another software program known as UNIFORM which reprocesses the data to eliminate discrepancies associated with the data. The program searches the database and flags parameters for which no value has been reported or for which the reported value does not conform to the format that the computer expects. Acidity, for example, is usually reported in terms of ppm CaCO_3 equivalent. It is however., often reported in negative numbers, as zero, with a value of -1 or it may not be reported at all. In order to process the data, the computer must be informed of such a situation. Although acidity is used as an illustrative case in point, all data parameters have designated reporting criteria that must be made uniform if the database is to be useable. This is especially true in the case of missing data or that data reported as zero.

After UPDATE, three uniform datasets are created as a working database. Once the working databases are generated, these files may be searched for any desired information by the COMMAND EXECUTION software. This software operates from command level which means that it is itself not a single computer program but rather a menu of programs that are available to be selected by the computer system depending upon the kind and amount of information, the specific data treatment and the type of output requested by the user.

The first program called by the EXEC is ENTRY which is a series of prompts by CRIS asking the user how definitive the requested information is to be. The computer asks whether the user wants a single permit or a list of permits, the kind of mine, the geographic location, the coal seam or seams and whether overburden or water information (or both) is desired. The final product of ENTRY is a temporary working file containing only those permits request by the user or those permits that contain the specific parameter(s) requested by the user.

Upon returning to the command level, the user can then select an output format including:

1. tabular
2. modified tabular
3. advanced systems with no tabular output or
4. advanced systems with tabular output

The TABULAR output will generate listings or tables of the requested information. The MODIFIED TABULAR produces the same listings or tables but generates them on a high quality printer. The ADVANCE SYSTEMS selections will result in a prompt which will ask for the specific request which is typed in by the user.

Possible selections include:

1. X-Y plots
2. X-Y-Z plots
3. histograms
4. pie Charts
5. choropleth maps (2D with plotted information)
6. block maps (3D with information plotted as blocks or)
7. surface maps (3D with information plotted as pyramids)

After selecting which of the above specific formats, the user then types in the specific information to be plotted or mapped.

Another possible request under ADVANCED SYSTEMS is statistical analysis. The program itself will provide four statistical options; two each for the water quality and overburden data. One can request the means of all water quality parameters for all samples from a specific site or one can request three month means (for seasonal effects) of all the water quality parameters for all samples from a specific site. With respect to the overburden data, the user can request two ratios. First the user can request a ratio of the thickness of a specific rock lithology to the total overburden thickness. The request will also provide a ratio of all compositional parameters of the selected rock type to the mean composition of the total overburden. The second ratio that can be requested is a ratio of the thickness of one rock type to that of a second rock type. The program will also provide a ratio of all compositional

parameters between the two rock types.

For more extensive statistical analyses such as multivariate analyses, the file can be transferred with the specific request(s) to a second computer that has the software to ready the data for transfer to AMD processing by statistical software on yet another system .

Perhaps the most powerful feature of CRIS is the mapping capability. Any parameter, combination of parameters or statistic can be plotted in either two or three dimension on a state-wide, county, quadrangle or watershed basis or within a specified radius from any geographic location. Such maps can of course be used in overlay with geologic, topographic or demographic maps.

CONCLUSION

A database management software package, CRIS, has been designed, generated and is now being tested for the compilation, searching and analysis of a large AMD database with a variety of tabular and graphic output options. The software will provide information to researchers, the coal industry and regulatory agencies to be used in the solution of AMD problems, feasibility studies and impact assessment.