✓
MASTER

DP-MS-76-75

CONF-761051--1

DATA MANAGEMENT AND ANALYSIS SYSTEMS FOR
LARGE-SCALE HYDROGEOCHEMICAL RECONNAISSANCE

by

R. B. Ferguson, J. H. Maddox, and H. F. Wren

Savannah River Laboratory
E. I. du Pont de Nemours and Co.
Aiken, South Carolina

A paper prepared for presentation to the
5th Geochautauqua meeting on "Computer
Software for the Geosciences" in Syracuse,
New York, on October 29-30, 1976

This paper was prepared in connection with work under Contract No. AT(07-2)-1 with the U. S. Energy Research and Development Administration. By acceptance of this paper, the publisher and/or recipient acknowledges the U. S. Government's right to retain a nonexclusive, royalty-free license in and to any copyright covering this paper, along with the right to reproduce and to authorize others to reproduce all or part of the copyrighted paper.

NOTICE

This report was prepared as an account of work sponsored by the United States Government. Neither the United States nor the United States Energy Research and Development Administration, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness or usefulness of any information, apparatus, product or process disclosed, or represents that its use would not infringe privately owned rights.

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

leg

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency Thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

DATA MANAGEMENT AND ANALYSIS SYSTEMS FOR LARGE-SCALE HYDROGEOCHEMICAL RECONNAISSANCE*

by

R. B. Ferguson, J. H. Maddox, and H. F. Wren
Savannah River Laboratory
E. I. du Pont de Nemours and Co.
Aiken, SC 29801

ABSTRACT

The Savannah River Laboratory has developed a versatile, computerized data recording, processing, updating, and retrieval system for handling an expected 150 million bytes of hydrogeochemical data from 150,000 to 200,000 sample sites over the next four years. A sub-system accounts for the movements of samples from initial receipt through final storage. Approximately 6 million sample movements are expected. Two- and three-dimensional plots of sampled geographic areas showing concentrations and locations of individual chemical elements are displayed and reproduced photographically. Pattern recognition techniques enable multivariate data to be categorized into "clusters" which indicate sites favorable for uranium exploration.

* Work done under USERDA Contract No. AT(07-2)-1.

INTRODUCTION

The National Uranium Resource Evaluation (NURE) program was initiated by the United States Atomic Energy Commission (AEC) in the spring of 1973 and continues under the sponsorship of the United States Energy Research and Development Administration (USERDA). The NURE program is a comprehensive effort to evaluate domestic (U.S.A.) uranium resources and to identify areas favorable for uranium exploration. There are seven major areas within the NURE program (SLIDE 1):

- National aerial radiometric reconnaissance.
- National hydrogeochemical and stream sediment reconnaissance (HSSR).
- Surface geologic investigation projects.
- Geologic drilling projects.
- Remote sensing projects.
- Geophysical technology development.
- Low-grade (<500 ppm U) study projects.

The goal of the NURE HSSR (SLIDE 2) is to complete (as much as possible by 1980) a reconnaissance of the Nation's surface waters, ground water, and stream and lake sediments. The reconnaissance will cover regions selected according to geologic favorability to aid in identification of the best areas for uranium exploration.

In 1975, the Savannah River Laboratory (SRL) in Aiken, South Carolina, operated by the Du Pont Company for ERDA, accepted

responsibility for hydrogeochemical and stream sediment reconnaissance surveys of twenty-five states in the eastern United States. Oak Ridge Gaseous Diffusion Plant (ORGDP), Los Alamos Scientific Laboratory (LASL), and Lawrence Livermore Laboratory (LLL) have accepted responsibility for similar reconnaissance surveys in the rest of the continental United States including Alaska.

The areas of responsibility (SLIDE 3) are based on geologic provinces normalized to state boundaries. Assignments are based on geologic provinces because, within them, geochemical factors that control both occurrences of uranium deposits and hydrogeochemical expression of uranium should be similar. For administrative reasons, areas of responsibility are coincident with state boundaries. The significance of the distribution of uranium in natural waters and sediments will be assessed as an indicator of favorable areas for the discovery of uranium deposits.

The present paper describes the participation of SRL in the NURE program and reports on the data management and analysis systems developed to support this program.

COMPUTER HARDWARE

A simplified schematic view of SRL's computer hardware (SLIDE 4) shows an IBM System/360, Model 195 computer, as the primary computing system. The System/360, Model 195 central processing unit (CPU) has a main storage capacity of 3,145,728

(3KK) bytes. Machine cycle time is 54 nanoseconds with an access time of 778 nanoseconds. The computer system has a total of 2894 KK (~3 billion) bytes of fast, direct-access, disk storage. In addition, 12 tape drives are available with recording densities up to 6250 bits per inch (BPI).

Program input/output and execution can be done from any one of 46 cathode ray tube display terminals. The terminals are controlled by the JOSHUA operating system.*

A computer output microfilm unit (COM) is available to users desiring to prepare printings, graphs, charts, maps, and other graphic representations from the computer system. COM output may be in the form of microfiche, 16-mm microfilm, or 35-mm microfilm. The COM system allows up to 192 pages of print on each microfiche with output speed of up to 40,000 characters per second.

An IBM System/360, Model 30 computer plays an important supporting role. Its use may be summarized in three categories:

1. Support of business programs.
2. Supplementary unit record input/output (I/O) when on-line to the Model 195 computer.
3. Utility service such as tape printing and special form printings, such as NURE labels, etc.

* JOSHUA is an operating system that provides the data handling services required to accomplish extensive, iterative, scientific calculations on a large, semi-permanent data base.¹

A Digital Equipment Corporation PDP 11/45 central processing unit with 72K bytes of core, tape drive, and two fast, direct-access, disk drives is used to support a 36-in. x 48-in. electronic digitizer and a Vectorgraphics 3-dimensional cathode ray tube display terminal (VG CRT). The VG CRT has the capability of displaying 2- and 3-dimensional plots that can be rotated around any axis, scaled larger or smaller in any dimension, and translated in any direction. The electronic digitizer allows digitizing the latitude and longitude of sample sites to within $\pm 0.0003^\circ$. A map containing 10 sites can be digitized in 5 to 10 minutes.

NURE SOFTWARE

The software requirements for a large-scale hydrogeochemical reconnaissance has three major subdivisions (SLIDE 5):

1. Data Management
2. Sample Inventory and Control System
3. Data Analysis and Interpretation

I will briefly discuss our efforts to date in these three general areas. All programming has been done in FORTRAN IV by the Computer Applications Division of SRL.

NURE DATA MANAGEMENT SYSTEM (NDMS)

The NURE Data Management System (NDMS)² must meet certain unique requirements (SLIDE 6). It will have to handle over 150 million bytes of data in the next four years from 150,000 to

200,000 sample sites. As the program evolves, the type and volume of data maintained for a sample site will change. In order to be cost-effective and reliable, the NDMS must be able to cope with continually changing input formats and master record structures. The system must print reports on all or subsets of the data. The system must retrieve subsets of data and make them available to statistical programs for analysis.

The major data processing steps (SLIDE 7) required of the NURE data management system are done by two programs, UPDATE and QUERY. Input data both from cards and from measurement analysis programs are used by a program (UPDATE) to update an old master file and produce a new master file. Information is selected and extracted from the master file and reformatted by a model (QUERY) producing reports and records stored in JOSHUA data sets.

The UPDATE program maintains the NURE master tape file. This includes adding a new record to the file, adding data to, or correcting portions of an existing record, and deleting a complete record. For the NURE master file, a record contains all data pertaining to a survey site plus the data that result from the analysis of the samples taken at that site.

The function of the QUERY module is to extract data from the NURE master records and write the data in a form for processing by other modules (e.g., JOSHUA numerical analysis modules) and in reports. The QUERY module reads each record in the master

file to determine if it meets the selection criteria supplied by the user. Output from QUERY consists of printed reports and JOSHUA records containing data fields as specified by the user. These records serve as input to various analysis programs which are not part of the data management system.

The current version of QUERY limits the number of items defining the records to be retrieved. The items defining the retrieval of records are survey site, sample source, sample type, and type of measurement.

The UPDATE program is separated into two modules (SLIDE 8). The Input Scrub and Sort module scrubs the input for syntactical errors and builds a file of syntactically correct input for the UPDATE module. The UPDATE module reads the input and the old master file. The records from the master file are changed to reflect the input and then written to the new master file.

The UPDATE module provides the flexibility required to accommodate anticipated changes in input, file structure, and similar details of the NURE data management system. The key to this flexibility is to control such details through a set of systems tables defined by the user, which are stored in separately maintained JOSHUA files. Examples of a portion of a Scrub System Table are shown in SLIDE 9 and SLIDE 10.

SAMPLE INVENTORY AND CONTROL SYSTEM

The sample inventory and control system (SICS)³ (SLIDE 11) was developed to maintain information for all samples (and sample splits) and the associated data forms. This system accounts for the movements of samples from initial receipt onsite through final storage. Approximately six million sample movements are expected.

The system has a disk file, which contains data for currently active samples, and a tape file, which contains data for samples which are no longer active. The disk file is maintained by a "direct-access" data manager. The tape file is maintained in sequential order by sample site code. Priority scheduling of analysis, error messages, shipping lists, etc. are written from the disk file by a "Report" module.

DATA ANALYSIS SYSTEM

The Data Analysis System (SLIDE 12) has two major subdivisions:

1. *Data analysis* for pattern recognition
2. *Data graphics* for 2- and 3-dimensional plots

Because of the large volume of data to be interpreted, data analysis and graphics play a central role in the hydrogeochemical program.

SRL uses the data analysis pattern recognition program ARTHUR (SLIDE 13) written by Dr. B. R. Kowalski, Laboratory for Chemometrics, University of Washington, Seattle, Washington.⁴ ARTHUR is a versatile FORTRAN IV program for statistical and pattern recognition analysis of multivariate data. ARTHUR provides an easy-to-use, relatively computer-efficient collection of many statistical and pattern recognition tools for examining multi-dimensional data.

Pattern recognition (SLIDE 14), as embodied in ARTHUR, is a tool designed to aid in making a systematic "educated guess" analysis of multi-dimensional data. Pattern recognition techniques can be divided into two major groups — supervised and nonsupervised learning. Supervised learning programs have not been used to date at SRL because these programs require *a priori* knowledge of the classes or categories to which the data belong. Non-supervised learning programs (commonly, some type of cluster analysis) look for assignment of "natural" groupings or clusters of data in multi-dimensional space. When clusters are defined, the supervised learning programs can be used to determine what measurement or combination of measurements are causing the "natural" clusters to occur.

The nonsupervised approach is being applied to the elemental and chemical measurements from the SRL-NURE orientation water data (SLIDE 15). "Natural" clusters can be identified, and the

most important controlling measurements determined. Presently, two hierarchical clustering (also known as "Q-mode clustering") programs and a minimal spanning-tree cluster program are being used.

Linear and non-linear graphics play an equally important role for interpreting the data, as do the mathematical statistical techniques. A non-linear projection (SLIDE 16) from 9-space to 2-space of orientation water data clearly shows the distinct clustering of the data. The ability to project higher dimensional space to 2- or 3-space while retaining maximum information is an extremely powerful interpretation tool.

A linear symbol plot program (SLIDE 17) generates a map where the symbols increase in size and darkness as the concentration of an element increases. The symbols, which are similar to the Canadian geochemical plotting symbols,⁵ are drawn by computer on microfilm. They can be photographically enlarged as symbol overlays to standard 1° x 2° Army Map Service (AMS) sheets or geological map sheets.

Three-dimensional (latitude, longitude, and elemental concentration) plots (SLIDE 18) on computer-generated microfilm or 3-dimensional CRT display are very helpful in determining anomalous locations. Programs allow the reference point of view to rotate in both horizontal and vertical directions. In addition, the point of reference can move closer or farther from the plot. This allows "blow-up" features on isolated structures.

Additional graphic programs allow for cumulative frequency plots (SLIDE 19) of selected data points to be generated on computer-outputted microfilm. This eliminates slow hand-plotting with its many drawbacks. Cumulative frequency plots can be computer drawn in less than one second from data selected by the QUERY data management subprogram.

REFERENCES

1. H. C. Honeck. *The JOSHUA System*. USERDA Report DP-1380, Savannah River Laboratory, E. I. du Pont de Nemours and Co., Aiken, South Carolina (1975).
 2. J. H. Maddox, H. F. Wren, H. C. Honeck, C. R. Tharin, and M. D. Howard. *NURE Data Management System*. USERDA Report DP-1432, Savannah River Laboratory, E. I. du Pont de Nemours and Co., Aiken, South Carolina (1976).
 3. *Savannah River Laboratory Quarterly Report. Hydrogeochemical and Stream Sediment Reconnaissance - Eastern United States National Uranium Resource Evaluation Program*. Savannah River Laboratory, E. I. du Pont de Nemours and Co., Aiken, South Carolina.
<i>Quarter</i>	<i>SRL Doc. No.</i>	<i>ERDA-GJO Doc. No.*</i>
January-March 1975	DPST-73-138-1	GJBX-5(76)
April-June 1975	DPST-75-138-2	GJBX-6(76)
July-September 1975	DPST-75-138-3	GJBX-7(76)
October-December 1975	DPST-75-138-4	GJBX-8(76)
January-March 1976	DPST-76-138-1	GJBX-17(76)
April-June 1976	DPST-76-138-2	GJBX-27(76)
4. D. L. Duewer, J. R. Koskiner, and B. R. Kowalski. *Documentation for ARTHUR, Version 1-8-75: Chemometrics Society Report #2*. Laboratory for Chemometrics, Department of Chemistry, University of Washington, Seattle, Washington (1975).
 5. *National Geochemical Reconnaissance - Saskatchewan 1974*. Open File Report 266. Geological Survey of Canada, Ottawa.

* Order from: E. W. Grutt, Jr., Manager
 Energy Research and Development Administration
 Grand Junction Office
 P. O. Box 2567
 Grand Junction, CO 81501

SLIDE 1

MAJOR DIVISIONS OF NURE PROGRAM

- National Aerial Radiometric Reconnaissance
- National Hydrogeochemical and Stream Sediment Reconnaissance (HSSR)
- Surface Geologic Investigation Projects
- Geologic Drilling Projects
- Remote Sensing Projects
- Geophysical Technology Development
- Low-Grade (<500 ppm U) Study Projects

SLIDE 2

NURE HSSR GOAL

To complete (as much as possible by 1980) a reconnaissance of the nation's surface waters, ground water, and stream and lake sediments (in regions selected according to geologic favorability) to aid in the identification of the best areas for uranium exploration.

SLIDE 3

ASSIGNED AREAS FOR NURE PROGRAM

SLIDE 4

SRL COMPUTER HARDWARE CONFIGURATION SCHEMATIC

SLIDE 5

SOFTWARE MAJOR DIVISIONS

- Data Management System
- Sample Inventory and Control System (SICS)
- Data Analysis System

SLIDE 6

DATA MANAGEMENT SYSTEM REQUIREMENTS

- Store, Retrieve, and Analyze ~150 Million Bytes of Data
- Generate Reports and Tables on All or Subsets of Data
- Retrieve Subsets of Data for Statistical Programs
- Be Insensitive to Input and Output Changes

SLIDE 7

NURE DATA MANAGEMENT SYSTEM: PROCESSING OF HYDROGEOCHEMICAL DATA

SLIDE 8

UPDATE FUNCTION FOR NDMS

SLIDE 9

PORTION OF A SCRUB SYSTEM TABLE

<u>Card Column</u>	<u>Scrub Option</u>
1	1
2	1
3	1
4	1
5	4
6	5
7	5
8	9
9	0
10	5
11	6
12	4

SLIDE 10

DEFINITIONS OF SCRUB OPTIONS

<u>Option</u>	<u>Character String</u>
0	No Scrub
1	A-Z
2	A-Z or Blank
3	0
4	0-9
5	0-9 or Blank
6	0-9, Blank or -
7	0-9, Blank or A-Z
8	1-9
9	Blank
10	Decimal or Blank
11	Blank, 0, 1, or 2
12	Decimal
13	+, -, or Blank

SLIDE 11

SRL-NURE SAMPLE INVENTORY AND CONTROL SYSTEM

SLIDE 12

DATA ANALYSIS SYSTEM COMPONENTS

- Data Analysis
 - Common Statistical Programs
 - Pattern Recognition Programs
- Data Graphics
 - Symbol Plots
 - Three-Dimensional Plots
 - Cumulative Frequency Curves
 - Non-Linear Maps
 - Factor Maps (Eigenvector Plots)

SLIDE 13

SRL PATTERN RECOGNITION CAPABILITIES

- Computer Code - "ARTHUR"
- Large, versatile, general-purpose, high-speed
- Many generalized pattern recognition subroutines
 - Pre-processing
 - Displays
 - Supervised and unsupervised classifications
- Batch Job Mode (at this time)

13

SLIDE 14

OBJECTIVE OF PATTERN RECOGNITION

"... to find and/or predict a property which is not directly measurable but exists by some unknown relationship..."

14

SLIDE 15

SLIDE 16

NON LINEAR DISPLAY of Penn. Water DATA

NLM
VARIABLE
2

NLM VARIABLE 1

1. *Staphylococcus aureus* (100%)

COMPUTER-GENERATED LINEAR
SYMBOL PLOT

SLIDE 18

COMPUTER-GENERATED URANIUM DISTRIBUTION PLOT

SLIDE 19

COMPUTER-GENERATED CUMULATIVE FREQUENCY PLOT