RPI-FWI-72-16A

72-16A

asis

AMERICAN SOCIETY FOR INFORMATION SCIENCE

1972 ANNUAL MEETING

October 23-26

Shoreham Hotel, Washington, D.C.

CONTRIBUTED PAPERS

AUTHOR FORUMS

Wednesday Evening, October 25, 1972

AMERICAN SOCIETY FOR INFORMATION SCIENCE
1140 Connecticut Avenue, N.W., Suite 804
Washington, D. C. 20036

Contribution No. 35 from the Eastern Deciduous Forest Biome, US-IBP

FIND

Freshwater Institute Numeric Database

Robert Kohberger
John Fisher
John Wilkinson

Rensselaer Polytechnic Institute
August 1972

School of Management

Research supported in part by the Eastern Deciduous Forest Biome, U.S. - International Biological Program, funded by the National Science Foundation under Interagency Agreement AG-199,40-193-69, with the Atomic Energy Commission-Oak Ridge National Laboratory.

Introduction

FIND - Freshwater Institute Numeric Database - is an information system for scientific data management. The system is designed to aid investigators participating in a long-range research project at Lake George, New York. The objective of the research is to study the aquatic ecosystem of a soft-water oligotrophic lake. Investigators on the project include biologists, climatoligists, hydrologists, environmental engineers, system engineers, statisticians, mathematicians, social scientists, geologists, chemists, and computer scientists. The information system permits rapid interrogation concerning what data are available, allows easily accessible storage of data, facilitates interchange of data among investigators, and makes possible ecosystem simulation and analysis using data from multiple sources. All of these functions are important when dealing with such a highly complex, interdisciplinary research project.

Data that are stored and retrieved under the FIND system include many different types of information. The primary types of information are measurements of the different processes occurring in the aquatic ecosystem. The total system will also include other types of information such as certain census data and land-use data of the surrounding watershed.

The FIND system has been designed to operate under ALPHA, a telecommunications procedure for the IBM-360. Through telecommunications, investigators may easily access the database from sites located off the Rensselaer campus, which is important as the investigation involves researchers at several neighboring institutions.

There are two main modules to the total information system. The first module, ADLIB, stores and retrieves abstracts containing descriptions of the data sets that are stored in the database. The second module, FIND, contains the programs that manipulate the numerical data.

Data Hierachy

A record of data stored in the computer consists of the actual data and identifiers giving supplemental information about the data. We have called these identifiers classification states. These classification states enumerate the how, what, when, and where identifications of data.

These classifications may be made by either direct recording or by locational significance. While each data set will usually have different.

classification states, the following examples are typical.

Examples:

1. Diatoms have been counted (3 species) at different stations (2), depths (4), and time of the year. Classification states are:

   a) Species - coded 1,2,3, for instance

   b) Station - coded 1,2

   c) Depths - actual depth

   d) Data - actual date

   The record would be

```
      Station, Depth, Date, Species, Data-Count
i.e.     1     3.0   6/17/69   1        100
```

2. Locational significance could be used by having all three species on one record. The first count would be for species 1, the second for species 2, and so on. The record would now be

```
      Station, Depth, Date, Data Count, Data Count, Data Count
                            Species 1   Species 2   Species 3
i.e.    1       3.0  6/17/69  100           50          75
```

In addition there are two more classification states that will always be present. These are

(i) password -- identifies the originator of the data, and

(ii) identification (ID) code -- a numeric code that identifies the specific data.

For example,

```
   Password - SMITH
   ID code  - 1 hourly incoming solar radiation
              2 air temperature readings
```

The password and ID code, in addition to providing information, control access to data files. The password and ID code must be correct before the data files may be read.

Classification states are quite important because they are information used by the retrieval program. The more detailed the classification states are, the finer the resolution that is possible in the retrieval program. If it is desired to have all the solar radiation values for the months of June and July between the hours of 1100 and 1300 (on 24 hour clock) the classification states associated with solar radiation must include time of day and date.

The FIND programs allow a maximum of ten classification states for each data set, which experience indicates is reasonable for aquatic data. These states are chosen in consultation with the investigator responsible for the data and a representative of the data processing group associated with the project. This representative has knowledge of the needs of all investigators so that the entire project is able to influence the selection of classification states.

ADLIB

Users must first be able to interrogate the database to determine what data are available. A file of abstracts is stored which may be searched and printed. The abstract is shown in Figure 1. Included in this information are the classification states of the data. With this knowledge, the user may enter the FIND programs to retrieve the desired data.

The program is designed to operate in ALPHA conversational mode. Briefly, through the use of command words the user requests a search of the abstract file. When the search is completed the program returns the abstract numbers of the abstracts found. The user may then request the printing of the retrieved abstracts.

Table 1 presents the command words of the program along with a description of their use. It should be noted that the INPUT command word is protected with a password so that unauthorized users may not enter abstracts.

FIND

The prime function of the FIND programs is to produce data sets that are of interest to the investigator. These data sets may be used directly as input to analysis programs. If they are not to be used immediately, they may be transferred to punched cards. The programs are designed to run with telecommunications procedures. The investigator at a teletype or IBM 2741 may execute the programs, input the information needed, and direct the output. The printing of large data sets and punching of cards is done at the Rensselaer Computing Center. This output will be delivered by courier to the investigators.

The following table presents a summary of the FIND programs and a short description of their use.

Figure 1
Data Set Abstract

-Sub Cat-

____Decomposition

____Hydrology

____Meterology

____Chemistry

____Terrestrial

____Pri. Productivity

____Sec. Productivity

____Modeling

-Author- Data Originator ____________

____

-Title- ____________

____

____

-Avail- To whom data are available:

Open or Restricted ____________

____

-Keywords- ____________

____

____

____

-ABSTRACT - Description of data

-Date- Dates & Places Data Collected ____________

____

-Taxon- If applicable, species studied ____________

____

Table 1
ADLIB Command Words

Word	Description
HELP	Describes all command words and the format and inputs needed for their use.
PRINT	Will print selected data set abstracts or if desired the entire file.
INPUT	Program recognizes that data set abstracts are to be added to the database.
SEARCH	Program accepts categories and words for which it must search the data set abstracts. The categories available for search are: 1. author 2. subject category 3. keywords
END	Program recognizes the user has completed his job and wishes to stop program execution.
FIND	Links the ADLIB program with the FIND programs to allow for retrieval of data.

Table 2

Summary of FIND Programs

Program	Description
NEW	Adds new data to the database. This program inserts data in the correct position sorted by classification states.
PURGE	Removes data from the database and consolidates the database.
SEEK	Searches the database for specified records and stores the retrieved records on a temporary file.
LIST	Produces a listing of SEEK's temporary file.
PUNCH	Produces punched card output of SEEK's temporary file.

Maintenance Routines - NEW, PURGE

These routines load new data to the database and change, add, and delete existing records. Their function is to keep the database updated with the most current data available from the investigators. Data are added to the database sequentially by classification states. For example,

1. A new set of data, SMITH-2, would be added after SMITH-1.
2. Suppose SMITH-1 has classification states only as station and date. If the stations are currently coded as 1 and 7, and data are added which are coded as SMITH 1, station 4, NEW will then add station 4 between 1 and 7.

Retrieval Routine - SEEK

The retrieval section of the database consists primarily of the program SEEK. SEEK will locate on a given file those data records that correspond to the classification states desired. These records are then stored on a numbered work file. This numbered work file is also available for retrieval by the SEEK program. Several searches may be made with the results stored in several numbered work spaces.

The usual method of locating specific data from files is the read/test approach. A special feature, file scan, hardwired into 2314 and some 2311 disk drives was utilized rather than the read/test method. With file-scan the input/output channel locates and retrieves the data while the CPU is freed for other jobs. This feature is not supported by IBM and had to be developed specifically for this application; however, the developmental time is more than justified by the savings in CPU time during execution.

These work spaces may be operated on by the use of "and/on" statements. Such statements have the effect of merging files with similar classification states. For example,

Suppose there are two data sets stored on the database with the following classification states

STROSS-1	Station	Date	Depth	Time Incubation	Data
	1,4,7	1969-1971	0.5-15.0	on 24 hour clock	PMax

SMITH-1	Date	Time	Data
	1969-1972	on 24 hour clock	total incoming solar radiation

The object is to match the PMax values with the observed solar radiation values. The steps in retrieval are as follows:

1. Retrieve STROSS-1 Station 1; assign to work file 1
2. Retrieve STROSS-1 Station 4; assign to work file 2
3. Retrieve STROSS-1 Station 7; assign to work file 3
4. Retrieve SMITH-1; assign to work file 4
5. "AND" work file 1 and 4, By Time; assign to work file 1
6. "AND" work file 2 and 4, By Time; assign to work file 2
7. "AND" work file 3 and 4, By Time; assign to work file 3
8. "OR" work file 1,2,3; assign to 1

Step:

5. Work file 1 now contains PMax values at station 1 (all times and depths) and solar radiation values matched by incubation time (STROSS-1) and time (SMITH-1).

6. Work file 2 now contains PMax values at Station 4 matched by time with solar radiation.

7. Work file 3 now contains PMax values at Station 7 matched

by time with solar radiation.

8. Combines the matched sets to finally obtain on one file: PMax values matched with solar radiation for stations 1, 4, and 7.

With each retrieval, a status message informs the user of the number of records retrieved and where the retrieved data are stored.

Summary

The information system presented permits investigators at Lake George, through ADLIB, a ready means of determining what data is available for their use. The FIND programs permit a rapid retrieval of the desired data. The total information system is of great aid to all investigators who must examine data from multiple sources.