UCID-18752
ASCOT/80/6

MASTER

THE ASCOT DATA BASE MANAGEMENT SYSTEM

J. Barbieri
R. Nyholm
C. Castro
K. Hill

Lawrence Livermore Laboratory
Livermore, California

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

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.

The ASCOT Data Base Management System

J. Barbieri
R. Nyholm
C. Castro
K. Hill

Lawrence Livermore National Laboratory, University of California
Livermore, California 94550

July 1980

DISCLAIMER
This book 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.

ABSTRACT

The ASCOT data base management system is designed to handle the data produced by both the experimental and theoretical efforts of the DOE Atmospheric Studies in Complex Terrain (ASCOT) project. The data base envisioned is hierarchically structured, sparse, and compact. Information concerning any given data file is stored in a directory file. The data base management system uses a relational data management approach. Presently three management schema are being developed for use with the data base.

EB
DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

1. Overview

The Department of Energy is funding the ASCOT program to study atmospheric transport and diffusion properties in a complex terrain.¹ The ASCOT program is a multiple DOE funded laboratory and university effort with approximately a dozen separate groups presently participating in field and theoretical studies related to drainage winds. The geographical area in which the study is being performed is the Anderson Creek area of the Geysers Area in Northern California.

Participants in the ASCOT program have or will be performing a series of experiments over three year period. These experiments can be divided into two classes. They are:

Continuous monitoring

Intensive periods including tracer experiments

Each class of experiment involves a number of different experiments and measurement systems integrated into an experimental design to determine the characteristics of drainage flows and provide data for model validation purposes. The basic difference between continuous monitoring and intensive periods of data acquisition involves the timing and magnitude of the experiments. Continuous monitoring involves a limited number of measurement systems data consistent with estimating long term drainage wind identification and characteristics. During an intensive period many additional measurement system are deployed to investigate specific characteristics of drainage winds and their behavior during a diurnal cycle.

There are a number of measurement systems active in the Geysers area. They include:

- wind component measurement systems
- radiosone & tethersome measurement systems
- acoustic sounding systems
- doupler radar systems
- meteorological tower measuring systems
- lasar anemometer system

Within each of the above measurement systems there are a number of variables which are monitored during the duration of the experiment. These include wind speed and direction, relative humidity, and barometric pressure.

The experimental measurement design of the ASCOT project presents unique problems for the logical management of information. For example the present sequence of experiments in the Geysers area i.e. the span is four years (data will be collected in the Geysers area for four years). After the drainage wind study in the Geysers area, the ASCOT program will likely move to another area and study this or other complex terrain problems. Other data bases will be developed at that time to suit the needs of the new study.

In any particular measurement session data might well be collected in five minute intervals (or less) hence the resolution of the data base could be in the minute or second category. The data base must be able to handle three years of data in which the separation (resolution) between records is very short. For example, consider the continuous monitoring of wind speed, wind direction and temperature at five locations on a meteorological tower. If one record is measured every five minutes for a period of a year, there will be 2, 102, 400 data points. If the same data is collected for a two week experimental session however there will be only 80, 640 data points, or 3.84% of the number of points for the entire year. This example is obviously an extreme, however it does illustrate one of the intrinsic properties of this type of data base. The ASCOT data base has the property of being sparse.

The number of items collected by the ASCOT project will be substantial, however it is highly unlikely that the experimentalist, theorist, or modeler will need to access each individual item. In fact, it is unlikely that a given record need be directly accessed on a routine basis. What will be of interest to the individual researcher will be groups of records, individual files and/or groups of files. Statistical analysis, data reduction, and analysis is expected to involve hierarchical grouping of data. In order to access groups of data it is necessary to uniquely identify the properties of that group. In the ASCOT project the initial level of access is currently the file. There are a number of reasons for choosing the file as the initial level of access. First, it represents the natural category or grouping of any of the experimental efforts. Second, they can be uniquely identified by the experimental effort. Third, the properties of the file are readily available-if the experimental design is good. The identifying characteristics of the file is the first record in any file.

A data base file for the ASCOT project has an identifying record called the header record and an arbitrary number of data records. If an individual researcher needs a given file which satisfies or partially satisfies his requirements, he need only access the header record to determine whether a particular file potentially satisfies his needs. If a number of files satisfy the researcher's needs, then groups of files can be combined and processed. Statistical analysis, data reduction etc. (statistical analysis), is beyond the scope of the data management system.

2. The ASCOT Data Base

The ASCOT data management system is divided into two distinct operational entities. The first is data acquisition, the second is data retrieval. The task of the data acquisition system is to capture incoming data files. There are a number of steps involved in this process which are described in a later section. The data retrieval system extracts those files which satisfy a file identification search constraints submitted by an individual user.

A data management system operates on a data base. The ASCOT data base consists of individual data files plus the directory files. A directory file contains an identifying record from each data file in the data base. The data files are themselves grouped into logical groups called directories. For the ASCOT project the directories are defined by the measurement type and the date. For example, measurement type can fall into a number of broad categories, e.g.

<u>Measurement Type</u>	<u>Mnemonic</u>
Met-Tower data	:METOWER
Met-Three component	:MET3C
Met-Acoustic sounders	:METAS
Tracer-tower	:TTOWER
.	.
.	.
.	.

The date components consist of the year, -month component of the date. For example, suppose a data file is received at LLNL in which meteoroglogical tower data were collected in August 1979, that data file would be stored in a directory which has the label

METOWER:79AUG

The concept of directories is completely arbitrary. It allows an initial categorization of each data file without using a data retrieval system.

A data base constructed in this manner is compact in the sense that there exist no empty locations.

The directory file is a file in which a record is added each time an incoming data file is added to the data base. Each record contains enough information to uniquely identify the incoming data file and also describe its characteristics. Each identifying record contains the following information:

- a record identifying number
- date the incoming data file entered the data base
- collecting organization

- incoming data tape I.D. number
- file name
- length of data file
- initial date/time of the experiment
- final date/time of the experiment
- site identifier
- longitude
- latitude
- elevation
- measurement (sample) type
- key-word identifiers (five allowed)
- all variables measured

Obviously the length of any particular record is arbitrary. It is dependent only on the properties of the data file. For each data file in the data base there exists a corresponding record in the directory file.

The data management system is designed to initially interact with the directory file to determine which data files, if any, satisfies the requestors constraints. In a sense the directory file is a catalog which identifies each data file in the data base.

Once the directory file is queried to locate files which satisfy certain global constraints, those files can be considered a data base at a finer level of resolution. By structuring the data within a given file or a combination of files, it is possible to apply further constraints to reduce the number of files or records within a file to a desired number. This concept produces data base that is hierarchical in structure. The data base therefore can be viewed differently at different levels of resolution.

Descriptions of data bases originally evolved in the context of business or commercial applications in which the properties of a data base were qualitatively different than the properties of a scientific data base. Actually, the ASCOT data base is qualitatively different than most scientific data bases. A property which is common in most data bases is an assumed similarity between records within a given

file or between files within the data base. The ASCOT base does not have this luxury. The data base is a reflection of the ASCOT project itself, that is, a number of contributors, each supplying a unique set of data based upon their particular involvement in the program; hence it would be unreasonable to insist on apriori constraints determined by the desire to implement a particular data basing/data management technique. In fact, the only constraint imposed on the researcher is that the particular data file be unique and well defined with respect to the variables measured — that is, the researcher must be able to define what was measured, when was it measured, where was it measured, and by whom it was measured (contributing organization).

3. Data Capture System

Data capture or alternatively, data acquisition is the process of capturing raw data from either a field experiment, or a numerical calculation and archiving it in the data base. Much of the process is mechanical and manifestly uninteresting. It must also be stated that the data capture process is the most critical process in the entire data management system. If the incoming data are ill defined or incomplete, then even the most esoteric data management system will be useless.

The ASCOT data processing procedures subject an incoming data file to a number of processes before that file becomes a member of the data base. The data file is considered to be "in" the data base when and only when its header record is stored in the directory file. The acquisition of data occurs in four major steps.

The computer system on which the data files are archived is the CDC-7600 with its associated mass storage devices. A word on a 7600 is made up of 60 bits, hence all incoming data must, ultimately be converted to this format.

FIGURE 1. ASCOT Data Archive Procedure.

The first step in the acquisition process is the conversion the incoming data a disc file on the CDC-7600. Usually this process involves reading an incoming tape onto one of the CDC-7600 disks and then storing the file on a mass storage device. This file has an initial character (T) followed by nine identifying characters: Tnnnnnnnnn.

The data file is usually in some format other than the CDC-7600 format. Also, the data are usually blocked. Before the data can be manipulated into a standard form for the CDC-7600 it must be unblocked and converted to 6 Bit ASCII format. Since data has been received at LLNL in a vast number of arbitrary formats and an equally large number of blocking factors, this conversion task is different for each contributing laboratory and there is no one procedure which can accomplish the conversion (there are a suite of available conversion techniques). Experience has shown that in some instances, the documentation received with the incoming data files may have little or no connection with the actual structure of the data file. The task of converting one data file into an acceptable format for processing can become a extremely complex and potentially frustrating process.

Once the data file has been converted into a machine readable format, it is converted to binary format. At this time, all variables are converted to a standard set of units. This process is accomplished by a procedure called PRQ. All variables in any data file are converted to MKS units, hence a description of the variables with their appropriate units is necessary to perform the conversion. The output file from PRQ has a P as the initial character: Pnnnnnnnnnn.

The data file is now in a format which is acceptable for data basing, however there is no identifying header record. The procedure PREP is designed to add this record to the head of the processed data file. In order to add this record to the data file the ASCOT librarian usually needs added documentation from the contributing agency. All files processed by PREP have the initial character D attached to the file name: Dnnnnnnnnnn. It is this file which will be archived in the data base.

The final step in the data capture procedure is the addition of the header record to the directory file. This process is performed by a procedure called ASTORE. This procedure also instructs the ASCOT librarian to store the data file in an appropriate directory. ASTORE also prints a description of the data file which is to be stored in the ASCOT librarians file. The name of the file which is archived contains information which can be used to identify the agency responsible for gathering the data and also the location at which the data were collected. From the description of the directory name it is possible to determine the type of measurement and the date the experiment was performed. Hence an internal structure of the data base is

SAMPLE TYPE: YEAR/MONTH
AGENCY1: LOC1
AGENCY2: LOC2
.
.
.
AGENCYm: LOCm.

For example, suppose that three agencies, Lawrence Livermore National Laboratory (LLNL), Argonne National Laboratory (ANL), and Los Alamos National Laboratory (LAS) collected data at Anderson Valley and High Creek. Further assume that the measurements made were from meteorological tower and that the data were collected in August 1979.

METOWER: 79AUG
LLNL:ANDVAL:File 1
ANL:HICRK:File 2
LAS:ANDVAL:File 3

The above description of the ASCOT data capture process does not fully indicate the amount of effort involved in guaranteeing that the data submitted by the various

contributors are not lost or inadvertently modified. As a general principle imposed on the LLNL data management effort, no unnecessary modification of a given data file internal structure is performed. If a contributing agency wishes to delete, modify, or otherwise update the information in a particular file, they are required to perform those modifications at their respective laboratories and resubmit the file. The staff of the ASCOT data management group cannot determine the accuracy of another agencies data.

4. Data Retrieval/Management Systems

The information stored in the data base is useless if it cannot be retrieved in a reasonable amount of time. The process of retrieving data is dependent upon the structure of the data base - initially, the directory file and the retrieval procedures. A given data base management system is actually only a part of an entire process.²

FIGURE 2. Data Base Management System Interfaces

The user has an interface which issues instructions to the data management system, while the data management system allows access to the physical record—that is, the actual records or files stored on the computer. There are three major data base management techniques currently in use. They are:

1. The Relational technique
2. The Hierarchical technique
3. The Network technique

The relational approach involves the creation of relations between various entries at some level of resolution. For example, suppose that data from a set of experiments can be grouped into two tables — S: sample type, and L - location.

S:

S#	S-type	File #
S ₁	AIRQ	file #1
S ₂	MET3C	file #2
S ₃	MET3C	file #3
S ₄	STACK	file #4
.	.	.
.	.	.

L:

L#	Location	File #
L ₁	Anderson Valley	file #2
L ₂	Plant #2	file #1
L ₃	High Creek	file #3
L ₄	Plant #1	file #4
.	.	.
.	.	.

Obviously, tables can be created, combined, and various columns projected out; hence relations can be created and destroyed dynamically. The relational data base management was conceived initially by Codd² and expanded upon by Date³ in recent years. The relational schema is probably the simplest and most elegant of all the approaches. It is also the most flexible of all the approaches.

The hierarchical approach involves the manipulation of "trees" which connect similar entities to each other. For our previous example suppose we wish to connect all files which contain data collected at Anderson Creek:

FIGURE 3. Hierarchical Approach

To locate a particular file an particular branch of the tree is traversed to obtain the desired information. Traversing trees can be an extremely rapid operation if the tree exists; however, if the given tree structure does not exist, the tree must first be created — which is a process involving the reordering of the data base. There is also an apparent asymmetry which causes problems — that is, in the above example, location information is somehow "superior" to date data.

The Network approach solves some of shortcomings of both the relational and hierarchical approaches, however it is significantly more complex — since each item is connected to every other item through "links". The network approach is cumbersome to use because of the overhead needed to manage linkages.

For the ASCOT project, the relational approach was chosen to be the most desirable technique. There are a number of reasons for the choice. First, the structure of the data base is hierarchical, hence at any level of resolution, only a small portion of the data base need be manipulated, i.e. the data base is compact. Second,

the relational approach is extremely flexible in its ability to create, manipulate, join, project and destroy tables and columns within tables, without an excessive amount of overhead. Finally, there is, at LLNL a considerable amount of expertise in relational data base management techniques. There exists at LLNL a relational data base management technique called FRAMIS⁴ which is capable of handling large data bases with an arbitrary structure. It allows a large number of elegant operations on the data base. Two smaller data management schema are also being developed at LLNL. The first, RELATE, is an extremely fast technique that is being optimized for use on the DEC-11/780 VAX machine. This system will not allow all of the operations of FRAMIS, nor will it allow totally arbitrary data bases. It contains however all of the necessary operations without expensive operational overhead. RELATE is expected to be operational by August 80. The second data management system being developed is an extremely simple data retrieval system called RETRIEVE. It is designed to exploit the structure of the directory file only. It does not contain any elegant table manipulation techniques. RETRIEVE should be operational by May 80.

The data management scheme envisioned for the ASCOT project involves the capability of operating on the data base at different levels of resolution with three data management techniques:

FIGURE 4. Operational Data Management System.

From an operational standpoint there are three steps which are common to each of the above management systems. First, the appropriate tables must be created at some level of resolution. This will involve user participation. Second, an appropriate relational data base management system is selected and applied. Finally, the appropriate data/file is extracted from the data base. The ASCOT data management group at LLNL will interact with the data base in only the highest level. Based upon requests presented to the ASCOT data management group, the group will extract these files which satisfy the users constraints. The ASCOT librarian will then send those files to the user.

FIGURE 5. LLNL Data Retrieval System.

As previously described the pertinent information describing an particular data file is contained in the directory file. The directory file is continuously being updated, hence

it is not a static file. The first task is creating the appropriate tables. For the data base retrieval system RETRIEVE a program called BUILDTAB creates the necessary tables. Output from either process is a list of those files which satisfy the appropriate search constraints. The data base management system RELATE is expected to be significantly more flexible. Once the file list is produced, a procedure called GET will extract those files from the data base. The files will be copied to magnetic tape and sent to the requestor with the appropriate documentation. In the above description we have not considered the data base management system FRAMIS. FRAMIS will be invaluable for processing the large data sets within the data base.

5. Future Efforts

The data acquisition/data management described in this report is currently being implemented on the LLNL computer system. This computer facility cannot be accessed by other agencies. Because of the nature of some of the work at Livermore, researchers are not allowed access the computer facilities. All contributions and requests made of the ASCOT data base must necessarily go through the appropriate LLNL personnel connected with the ASCOT project.

It would be extremely useful if the entire ASCOT data base was directly accessible to the various agencies through remote terminal connections. Researchers at various agencies would have access to the data base and would be able to use the data management tools directly. Data could be processed interactively. At present, no such facility exists, however relatively small computers exist that are suitable machine for such activities. Only data retrieval and data analysis would be available. The respective agencies would not be allowed to enter data files to the data base. In order to insure consistency in the data base, the capturing of data files should continue to be controlled by ASCOT personnel at LLNL.

Presently the data base management system FRAMIS is being implemented on the DEC-11/780 VAX. Both RELATE and RETRIEVE are written for the DEC-11/780 VAX. If the researchers from various agencies have direct access to the data base then all three of the above techniques could be applied to specific needs.

Upon completion of activities in the Geysers area, the data base, together with the data management software is to be sent to a qualified outside institution. The institution will maintain the data base which will make data gathered in the Geysers area available to any qualified requestor.

6. Acknowledgement

The authors would like to thank Dr. Dickerson and Dr. Gudiksen for their input during the development of the ASCOT data management system. We would also like to thank Dr. Dickerson for his critical review of this document.

Vocabulary

In this report a number of commonly, but ambiguous terms are used. We shall define some of these terms.

1. An item is any unique value - whether measured or calculated - that is to be stored in a data base.
2. A record is a logical collection of items which are logically related to each other. For example, assume that at a given instant, a set of measurements are made: time (t), wind speed (ws), wind direction (wd) and temperature (T)-the record for this data contains four items:

t	ws	wd	T
---	----	----	---

For most records we shall need identifying items. In the above record, time (t) is an identifying item. A more completely identified record might contain the location (loc) and elevation (h). The record would contain three identifying items and three measured items:

t	loc	h	ws	wd	T
---	-----	---	----	----	---

3. A file is any collection of records which form a complete group. Using the above example, assume that the recording instruments which made the above measurements collected one record each hour for a week the file would contain 168 records or 1008 items. If the instrument remained in one location for the duration of the experiment 336 items would be repeated.
4. A data base is any combination of files, records or items which can be grouped into a complete set. In a sense a data base is the union of all information at

some level of definition. Using the above example, suppose that we wish to build a data base which contained one years data. The above data base would contain 52 files.

5. A table is any subset of a data base or a file. Assume that from the above data base we wish to look at only that information which contained time (t) and temperature (T). The appropriate table would contain 17472 items-8736 would be time (t) and 8736 temperature items.
6. A column is any given type of item; for example, temperature would be a column.

t ₁	T ₁
t ₂	T ₂
t ₃	T ₃
t _n	T _n

7. A data capture system is any system which captures incoming data in some coherent, consistent and complete manner.
8. A data management system is any system which is capable of querying, editing or altering the data base or some subset of a data base.

References

1. M.H.Dickerson and P.H. Gudikson, "ASCOT FY-1979 Progress Report," Lawrence Laboratory, Livermore, CA., UCRL-52479 (1980).
2. E.F. Codd, A Relational Model of Data for Large Shared Data Banks," CACM 13, #6 (1970).
3. C.J. Date, An Introduction to Data Base Systems, Addison Wesley Publishing Co., Menlo Park, CA, (1977).
4. E.W. Birss, S.E. Jones, D.R. Ries, J.W. Yeh; Scientific Data Base Management: Needs and a Prototype System, Lawrence Livermore Laboratory, Livermore, CA, UCRL-80146, (1977).

js/1088z/103z

This report was prepared as an account of work sponsored by the United States Government. Neither the United States nor the United States Department of Energy, 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.

Reference to a company or product name does not imply approval or recommendation of the product by the University of California or the U.S. Department of Energy to the exclusion of others that may be suitable.

Printed in the United States of America
Available from
National Technical Information Service
U.S. Department of Commerce
5285 Port Royal Road
Springfield, VA 22161
Price: Printed Copy \$: Microfiche \$3.50

<u>Page Range</u>	<u>Domestic Price</u>	<u>Page Range</u>	<u>Domestic Price</u>
001-025	\$ 5.00	326-250	\$18.00
026-050	6.00	351-375	19.00
051-075	7.00	376-400	20.00
076-100	8.00	401-425	21.00
101-125	9.00	426-450	22.00
126-150	10.00	451-475	23.00
151-175	11.00	476-500	24.00
176-200	12.00	501-525	25.00
201-225	13.00	526-550	26.00
226-250	14.00	551-575	27.00
251-275	15.00	576-600	28.00
276-300	16.00	601-up ¹	
301-325	17.00		

¹ Add 2.00 for each additional 25 page increment from 601 pages up.

Work performed under the auspices of the U.S. Department of Energy by the Lawrence Livermore Laboratory under Contract W-7405-Eng-48.