10/14

TECH MEMO

a working paper

System Development Corporation / 2500 Colorado Ave. / Santa Monica, California

TM- 1456/003/00

AUTHOR J. H. Bryant

TECHNICAL

RELEASE S. N. Cooney

for D. L. Drukey

DATE 18 Sep 63 PAGE 1 OF 12 PAGES

DEVELOPMENT AND MANAGEMENT OF A COMPUTER-CENTERED DATA BASE:

PROCEEDINGS OF THE SYMPOSIUM (10-11 June 1963)*

Part 3: AIDS/SAC Experience in Managing Data Base Operations

J. H. Bryant

*This symposium was conducted by the Command Systems Department's ARPA Command Research Project, in cooperation with the Advanced Research Projects Agency. This talk, one of the presentations at the symposium, is to be incorporated in a Technical Memorandum containing the complete proceedings.

AIDS EXPERIENCE IN MANAGING DATA BASE OPERATIONS

J. H. Bryant*

I intend to cover two subjects in this talk: a brief description of the AIDS/SAC System, and a discussion of some of the problems involved from the operations side of organizing and developing a computer-centered data base.

The AIDS/SAC System is a data processing complex designed to support SAC intelligence by providing data storage, retrieval and processing as required by the operational missions of the branches of the organization. The system provides data processing facilities for establishing, maintaining, and using a variety of data files for special analytical and processing purposes. This ~~brief description of the system will~~ emphasize the current software components, primarily the program system, that have been developed to partially satisfy the mission of AIDS/SAC. The manner in which the software system is designed to operate will provide, I think, a pretty good indication of how the employment or operational system functions. It can be described in terms of a few definitions and concepts of data records and files and in terms of basic operations that may be executed in the system.

In the AIDS/SAC filing concept, files are designed to contain data records about a particular subject or subject class. All records in a file conform to the same record definition or format within certain pre-established limitations. Data records are either of fixed or variable length. Variable-length records are composed of two types of data fields, designated as "fixed" and "periodic," periodic fields being those which may appear one or more times within the same data record, such as time 1, time 2, time 3, etc. A periodic group is a set of periodic fields designed for a particular type of datum. While there may be a number of groups within a particular record (this is a variable component) all periodic groups within that set must conform to the same definition. This concept is illustrated in Figure 1.

---

*Dr. Bryant is Assistant Manager for IBM's AIDS (Aerospace Intelligence Data System) activity with the Directorate of Intelligence at SAC Headquarters.

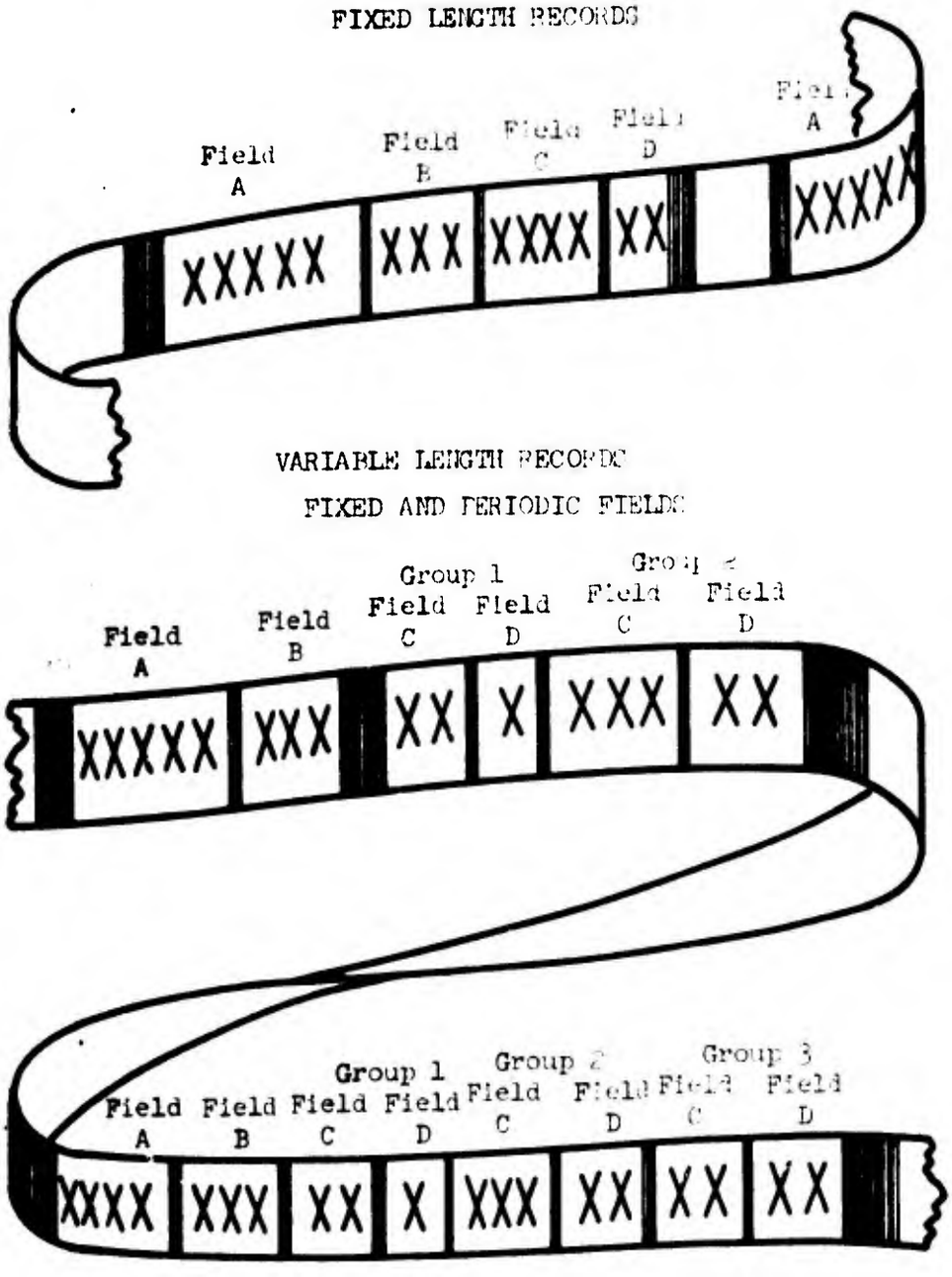


Figure 1. Illustration of the Two Basic Record Concepts Used in the AIDS/SAC System

The files themselves may be open-ended, i.e., extended over tapes with the addition of new records, or they may be of a fixed length. They may be maintained in a specified sort order or in order of data entry. Input records to a file may be prepared in either fixed or "comma" format as shown in Figure 2.

There are three basic job types that can be processed: file establishment and modification, file maintenance and update, and file query. All system operations are controlled by processing instructions prepared in the system macro language. The first statement in an instruction sequence contains the job control information and must specify one of the three job types, using these neumonic operations codes:

FILE ESTABLISHMENT AND MODIFICATION -- FMOD

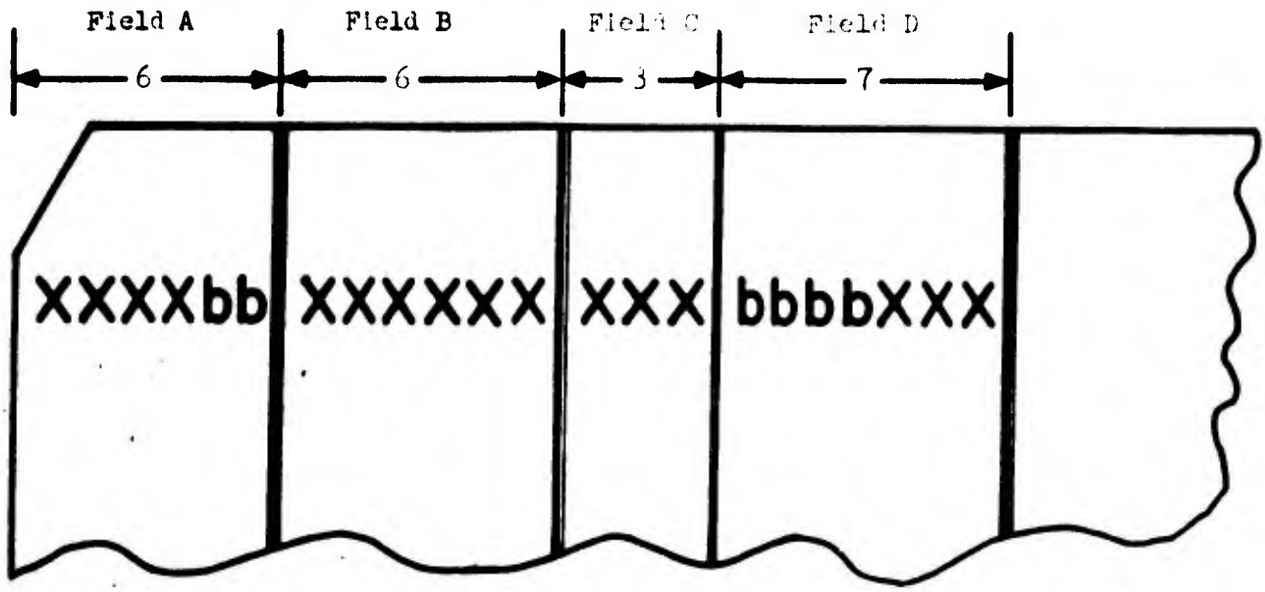
FILE MAINTENANCE AND UPDATE -- FMUP

FILE RETRIEVAL AND PROCESSING -- FQRY

Since all jobs are processed against a particular file in the system, the job must address a specific file using a neumonic file name. Jobs may be processed according to one of four priority conditions specified in the job control statement. Deferred priority jobs are processed in order of entry when no other jobs are in the system to be executed, while normal jobs are processed in order of entry, and have precedence over deferred jobs. This is presumably the normal mode of entry. Expedite jobs are processed before lower priority jobs in inverse order of entry. The highest priority jobs interrupt whatever is being executed on the machine and are operated on immediately.

Most of the processing instructions in a job sequence, after the control statement, address a specific data field in the file, which has been selected by neumonic field name; specify the logical connection between an instruction and the previous instructions in the sequence, and match conditions that must be satisfied and the value a field must contain. The types of operations that can be performed within each of the three job types are briefly summarized as follows. File modification jobs perform two major functions: they establish new files, or they modify the structure of existing files from specifications included with the job. File specifications define the file structures; establish symbolic addresses of the files in the data fields; specify the use of any special purpose conversion routines that will be executed on incoming records, such as conversion of meters to feet, central to eastern time; and so forth. They also specify the use of any table look-up operations used for the insertion of reference information into a data record (e.g., geographic coordinates when plain text place names are entered). The specifications entered with the file modification jobs also

FIXED FORMAT



COMMA FORMAT

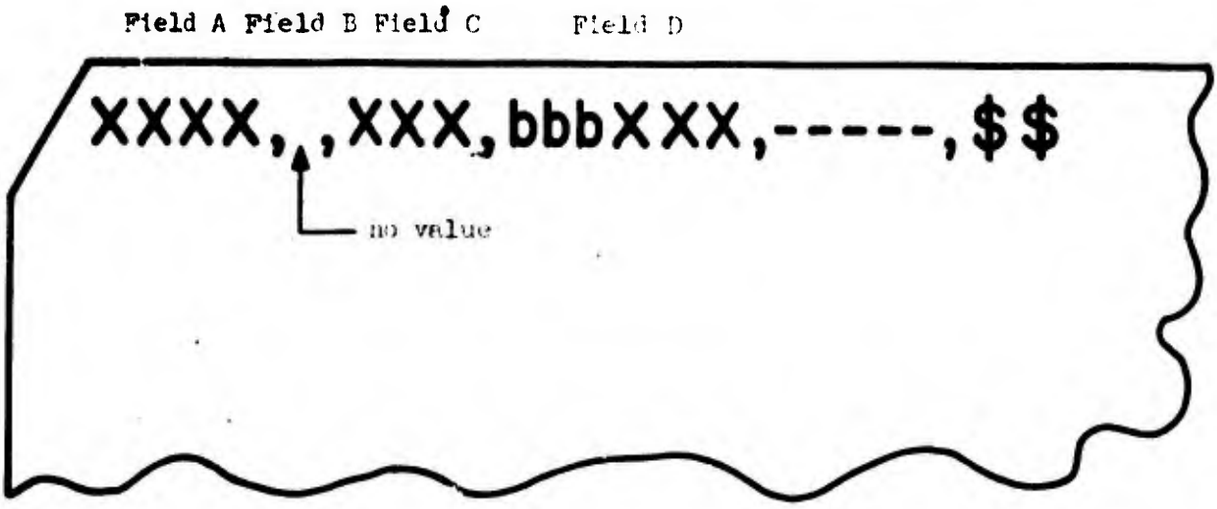


Figure 2. Illustration of the Concept of Fixed and "Comma" Formatted Input Records

contain tables used for conversion of coded field names and field entries to more meaningful output column headings and substitute phrases. For example, a data field that contains geographic place names may be designated as LOCA, for location, when instructions are prepared but may be printed out with the heading LOCATION or PLACE.

File maintenance jobs perform any combination of the following: they add new records to an existing file, add new data to existing records, change the contents of existing records, or delete existing records from the file. Records may be added in the front or the back of the file, or may be merged into a file that is kept in-sort. Out-of-sort inputs are automatically entered into the order of the file. If fields are added or changed within a record, or records are deleted, the exact record or group of records to be affected must be specified. This is done by specifying the values that the addressed fields must contain in "and/or strings," the condition to be satisfied, and the operation to be performed. A simple illustration of a complete file maintenance update job sequence follows.

```
FMUP, File Name, Priority, File Changes as of 63/06/10, $$

GET, FIELD A, EQUAL, X, $$

AND, FIELD B, GREQ, YY, $$

AND, FIELD B, LSEQ, ZM, $$

OR, FIELD A, EQUAL, X, Y, Z, $$

AND, FIELD B, NEQUAL, ZZ, $$

CHANGE, FIELD C, EQUAL, ALPHA, $$

END, $$
```

Figure 3. Illustration of Macro Instructions for Changing the Contents of an Existing Record or Group of Records

The first instruction line, FMUP, designates this as a file maintenance job addressed to a particular file, by name, with a specific priority processing condition and any identification that one wants to add to it. Following the job control, we begin by finding those records in which field A has a value equal to X, and field B has a value greater than YY and less than ZM; or field A is equal to X or Y or Z and field B is not equal to ZZ. Following the conditional portion of this sequence, we now say "Change field C so that it contains the value alpha" and we terminate the job.

File query jobs retrieve records from a master file according to the match logic specified, process the retrieved records according to instructions provided, and form output listings or tapes as specified.

Following is an illustration of a file query job.

```
FQRY, File Name, Priority, Unclassified Report, $$
BEGIN, FIELD A, EQUAL, X, Y, Z, $$
AND, FIELD B, NEQUAL, L, M, N, $$
----------ETC.----------
SORT, FIELD A, ASCD, $$
SORT, FIELD B, DSCD, $$
SUBTOT, FIELD C, FIELD B, $$
SUBTOT, FIELD C, FIELD A, $$
TOTAL, FIELD C,
OUT, BCDT, FIELD A, FIELD B,..., $$
END, $$
```

Figure 4. Illustration of Instructions for Retrieving

Again, the first instruction is job-control. Following the job control are the statements that define the record to be extracted out of the master file. This process is similar to the instructions in the file maintenance job. Following the extraction, we can now specify the type of processing that we want to be executed. In this case, we are sorting field A in ascending order and field B in descending order, and obtaining subtotals on the values in field C based on a break in field B, etc. Finally, we are writing out a BCD tape with the fields arranged in the order specified.

Within the context just described, systems operations are established to provide for data entry into the system directly from source reports or from pre-analyzed data from operating organizations. Some data are entered directly from paper tape, others are entered from card or tape records prepared elsewhere within the command, and other data are manually transcribed

and keypunched. Queries are generally received at the system central in narrative form and are coded into processing instructions by technicians.

Now let us discuss what is, I think, the major problem in management of a computer-centered data base: The control of operations and data errors.

The types of failures and errors that are of primary concern are those which result in permanent or temporary data loss, making the data unaccessible when needed and increasing the over-all operating expense because of re-run and recovery. Techniques used for control are not considered to be unique, except as they may apply to our particular system at SAC, but these techniques are often overlooked or underemphasized in systems planning. Control techniques fall into four interrelated categories keyed to the major component affected. These are: personnel or manual tasks, personnel organizations, file design, and software design. (Hardware is assumed to be a constant for any given operating system.) Since these four categories are considered to be interdependent I have not attempted to maintain a clear-cut distinction in the discussion that follows. Before discussing the control techniques that we have employed, let me first describe the processing environment at SAC.

The bulk of the data records of the current system files are either extracted directly or are derived from field reports. The majority of the reports that are processed directly by the system are forwarded by teletype on a pre-scheduled basis. One error that often occurs is the failure to receive all scheduled reports. Reports may contain data errors introduced during field preparation or during transmission. These of course, are only recognized when there is a violation of either some logical or arbitrary data specification.

Failures and errors originating outside the system, so to speak, can sometimes be monitored or detected but are less amenable to control techniques than are failures and errors within the system. Recovery is limited.

Reports that are handled within the system are subject to additional errors. The handling process involves an in-line production process in which incoming material is first sorted into files and operation categories, and is then routed to the next station, where it is edited and coded for card creation. Cards are punched and, subsequently, the files are updated. As mentioned already, the queries follow a similar data flow. At the first step, errors may occur from an inappropriate sort, which may result in an inadvertent loss of reports. At the next step, errors may occur in editing or coding either as the result of a failure to mark or extract all material, or as the result of inappropriate coding. Source material may be lost at any station or between stations. Keypunching errors occur at the next step; cards may be lost. At the computer processing step, errors may occur through the use of wrong system tapes or wrong data tapes for the particular job, or through failure to follow instructions or preestablished procedures.

In attempting to improve system performance with respect to the first control techniques category the category of personnel or manual tasks, written procedures for all operations were prepared and are continually modified. These include procedures to account for receipt of all scheduled reports. Imbedded within the routing process control procedures is a record system that provides for a throughput accounting of all documents during handling, through use of a batch-and-sequence-number control. Operations performance data are also collected at the same time to provide statistical data for manpower requirements, task allocation, scheduling and a general operations planning. Currently, procedures are prepared as a part of file operating specifications prior to the implementation or establishment of new files. This tends to increase the work load at the file design stage but minimizes errors at implementation and in subsequent operations. Sizable manpower effort is required to prepare operating documentation. For example, the written procedures in the four major categories for the system at SAC have amounted to more than 600 pages of detailed instructions. There have been 17 major revisions in a little over 18 months.

In addition to manual checks performed during data preparation, error detection programs, involving logical and consistency checks, and table look-up operations have been prepared for all inputs where both job processing instructions and the contents of the data records are checked. These programs are operated off-line on one of the peripheral computers following keypunching and job batching and prior to main frame processing. These have been designated as pre-editing operations. In the current system configuration, pre-editing and manual correction procedures have considerably reduced data errors and made the operation more efficient.

Since the data retrieval language provides the capability to examine in detail records stored in a master file, a post-edit is performed, following file updating operations, by a retrieval of records from the files that contain data entries that are not equal to legal values or not logically consistent This procedure of post-editing has also tended to reduce errors in the data base.

Most data file records in our system consist of at least two types of information: heading data, which identify the source reference of a particular record, and detailed, significant data. Procedures have been prepared for retrieving and crosschecking source information against incoming report-receipt logs to insure that all reports have been processed that should have been.

The personnel organization that operates the system has two groups responsible for operations and process control; one group is responsible for production coordination and control of job operations within the facility; the other, for file quality control. Members of the latter group, designated as file

analysts or file monitors, review the end result of all maintenance processing, to determine whether or not there have been inadvertent errors resulting in data loss; to take corrective action as necessary; and to make any modifications to data files or operating procedures required by changes in source data or by user requirements.

Skill requirements for the system operations personnel are somewhat unique because of the nature of the system. To provide for qualified personnel, a fairly extensive training program has been developed for base-level techniques training in system operation, and specific skill training in all system functions. The total course time is more than 7 weeks.

Some of the reports that are handled by the system are highly stylized or formatted. For one type of datum, programs have been prepared to update the file directly from paper tape inputs. Incoming data are formed into input records and extensively checked for logical and legal entry errors, and error corrections are attempted and made wherever possible. This operation, which represents about 30% of new record addition per day, bypasses all manual input operations and subsequently reduces the possibility of introduced error. In addition, manpower requirements have been reduced by 50% for that category of material. Within this process, records containing errors that are detected and that cannot be corrected by the program are saved, and the error fields are numbered and listed for correction, by number, with sufficient data environment to permit correction from a listing. The correction is made simply by reentering the identifying error number and the correct value.

I'd like to mention briefly two other factors that have contributed, I believe, to more efficient operation of the central data processing system at SAC. The first pertains to file design -- specifically the control of failures and certain types of errors. In the AIDS/SAC environment, several organizational entities may have a requirement for the maintenance and use of a single file; conversely, a single organization may require several files. Even though the files contain specific types of data, most data files contain certain data fields that are common to all (for example, message reference information, and date or geographic location fields). Since both file maintenance and file retrieval operations typically involve the writing of instructions that address particular fields within a record and specify exact values for those fields, it is desirable to standardize addressable names and codes, output headings, and substitute phrases throughout the files wherever this is possible. This helps to reduce the training or learning time, to provide more compatibility between users, and to avoid errors in preparation of processing instructions. It also simplifies the tasks involved in maintaining coding handbooks, since common fields, field values, etc., can be maintained in one place. (Incidentally, the latter isn't a very easy job. We have in our system one handbook of 2 volumes of about 200 pages of tables, which must be updated weekly.)

The last factor pertains to one aspect of file maintenance techniques. As mentioned earlier, a post-edit of master data files is performed, with the use of the retrieval program, following update operations The error detection query, as it is called, produces a list of file records that contain one or more errors. In most cases it is possible to correct the records from the context of the records or from filed source documents. With the first model of the program system, error records were corrected by, first, a deletion of the error records and, then, a reentry of the correct records. This involved preparation of two sets of processing instructions and, consequently, there was a higher probability of introduced error. Since it was otherwise necessary to locate the error records in the file to perform a deletion function, the program was modified to permit the selected records to be changed directly according to processing instructions in the form shown above. Thus, both functions were accomplished at one time with fewer processing instructions This technique, incidentally, has been used fairly extensively for other types of maintenance operations, particularly for files in which the normal maintenance consists of modifying the content of the existing records rather than of adding new ones.

The types of controls and the operations planning that I have just mentioned are probably standard in most production operations; obviously, they haven't eliminated all of our process problems, but they have been of some significant value to us. I'd like to reemphasize the major point when systems planning in data retrieval or computer-centered data bases takes place, it is essential to consider the employment system and how it is to be executed at the same time, and to give it as much emphasis as is given to software/hardware considerations.

Questions

D. L. Drukey. "Could you tell us something about the size of your file and the amount of input to the file and the number of queries per day that you get?"

Answer:

We have at the present time 13 files, varying anywhere from about 50 tapes per file to a quarter tape. That gives you some indication of size. Input volume for the largest file category amounts to about 3,000 records per day; for the smallest, about 100 records per day.

Question:

"What was the number of queries?"

Answer:

Number of queries? I think I already mentioned that most of the queries we

process are of the anticipated type. We process about 250 of these standing-type queries per month and right now, of course, we are having a much higher number because we are in system test configuration. I'd say in a typical operation it may not amount to more than 5 or 6 queries per day, of the spot-query type.

Question:

Licklider: "It seemed to me that in your discussion you were concerned with getting answers to questions, and were not very concerned with performing operations upon information restructuring and reinserting it in the file."

Answer:

That's correct. We do handle both types of operations within the system. There are finished data files that are derived from raw data, based on special analyses and computations. These analyses and computations are performed outside the environment of the program system so to speak. The system functions primarily for data storage and retrieval.