A Data Manager Looks at the Development of the Colorado Water Data Bank

by

Robert A. Longenbaugh and Norval E. McMillin

Department of Civil Engineering
Colorado State University
Ft. Collins, Colorado 80521

Paper presented at

"Symposium on the Management of
Data Elements in Information Processing"

National Bureau of Standards
Gaithersburg, Maryland

January 24, 1974

Colorado Water Data Bank Project
Technical Report 73-06

CEP73-74RAL-NM22

A Data Manager Looks at the Development of the Colorado Water Data Bank[1]

Robert A. Longenbaugh[2] and Norval E. McMillin[3]

Department of Civil Engineering
Colorado State University
Ft. Collins, Colorado 80521

The Colorado Water Data Bank Project is developing a central computerized data base for capturing, storing, and retrieving all types of water data collected in the State of Colorado. The three-year project includes development of all software programs, operational procedures, program documentation, and user manuals for capturing both current and historic records. The project is funded by the Colorado Division of Water Resources (DWR) and represents a cooperative venture between DWR and Colorado State University.

The first task for the Colorado Water Data Bank Project was to evaluate and choose a Data Base Management System (DBMS) to be used for the project. Following selection of the DBMS, a major effort was required to develop record formats and file structures which were compatible with the DBMS and would still provide efficient and economic storage with maximum retrieval flexibility. External programs, written in COBOL and FORTRAN, interface with the DBMS to perform editing and updating of data, as well as preparing sophisticated reports.

A system was developed for capturing, editing, reformatting, loading and retrieving the desired water data and is identified as the Colorado Water Data Bank System (CWDBS). A general flow diagram and a brief description of the system is presented.

Key words: Colorado; data standardization; Data Base Management System; MARS VI; water; Water Data Bank.

1. Introduction

The Colorado Water Data Bank Project was initiated July 1, 1972. The Project is funded entirely by the State of Colorado, and consists of a developmental and initial data capture phase to be completed in the first three years, followed by a continuation phase where additional data will be added from year to year as new records become available. The Project represents a cooperative endeavor between the Division of Water Resources (Colorado State Engineers Office) and Colorado State University.

---

[1]Authorization for publication granted by Colorado Division of Water Resources.

[2]Project Leader, Colorado Water Data Bank Project and Assistant Professor, Civil Engineering Department, Colorado State University, Ft. Collins, Colorado.

[3]ADP Programming Supervisor, Colorado Water Data Bank Project, Department of Civil Engineering Colorado State University, Ft. Collins, Colorado.

The Division of Water Resources (DWR) has a contract with Colorado State University to provide the computer facilities (CDC 6400) and to use its technical expertise to develop and implement the data bank system including programming, documentation and user procedures for capturing, storing and retrieving the data. The Division of Water Resources is responsible for capturing the historic and current records in a machine-readable format.

Colorado's rapidly increasing population and the corresponding development of competition between the agricultural, municipal, recreational, and industrial water users has magnified the water administration problems. Currently, the Division of Water Resources is required to administer both ground and surface water within the existing laws. Changes in administrative policies are continuously being evaluated to provide complete management of both ground and surface water supplies to minimize water shortages and to provide maximum beneficial use of Colorado's limited water resources. Different types of data are required for each administrative decision.

Recent legislation required that the Division of Water Resources also provide different types of data and administrative decisions to be incorporated into land use planning. A comprehensive land planning bill is being prepared in 1974 by the State Legislature and will require certain water-related data to be incorporated into comprehensive land-water use plans. Extensive use of the Water Data Bank is expected by federal, state, and local water administrators, as well as engineers, lawyers, economists, planners and the general public.

The primary reason for establishing the Colorado Water Data Bank was to provide at a central location all types of water data. The need for rapid administrative and management decisions requires that water data be readily accessible in a form which can be incorporated into simple or complex computer programs. The need to cross-reference different types of data also requires that the records be compatible and available at a central location.

Prior to establishment of the Data Bank Project, most of the data had been processed manually with data storage consisting of handwritten ledger books, keypunched cards, and in some cases, data stored on magnetic tape. For example, gaging station records were available from the U.S. Geological Survey's data bank in Washington D.C. and climatological data were available from U.S. Weather Bureau publications or on magnetic tape from the Weather Bureau Record Center at Asheville, North Carolina. Other examples include Colorado water well data stored on magnetic tape in the State Engineers Office and records for historic diversions, water rights, and descriptive data on dams which exist as typewritten or handwritten records in the State Engineers Office. The incompatibility of the data and the major time required to access and retrieve data from the many sources is quite apparent to those using the data.

2. Data Description

The nine different types of data to be incorporated into the data bank in the initial phase are illustrated in figure 1. These include information on climatology, gaging station records, ditch diversions, reservoirs, dams, water rights, wells, stock ponds, and eventually water quality. The lines connecting the circles in figure 1 indicate that cross-referencing between the connected types is needed. For example: In evaluating the adequacy of a water right to provide water for a proposed new subdivision, it is necessary to evaluate the water right as well as the historic amount of water which has been diverted. Development of cross-referencing identification numbers will be described later in this paper.

The oldest water right in Colorado dates back to 1852 and numeric records on the amount diverted have been kept since 1881. Table 1 indicates the magnitude of the different types of data which are to be entered into the Water Data Bank. The decision was made by the Division of Water Resources to place 30 years (1942-72) of historic diversion, reservoir, and climatological data into the Data Bank. To provide complete and accurate records it was necessary to develop the capability for capturing current data beginning with 1973. Only those types of data which had been recorded in the past were to be included in the data base; however, the system was to be capable of handling new types of data at a later date.

Table 1. Types of data to be initially placed in Colorado Water Data Bank with indication as to whether it is descriptive, numeric, or both.

Type of Data	Descriptive Data	Numeric Data
Water Rights	37,000 Records	--
Reservoirs	2,200	30 years historic monthly + current
Dams	2,500	--
Gaging Station	530	Daily values for entire record
Diversions	12,000	30 years of historic daily + current
Wells	75,000	--
Climatology	248	30 years of historic daily + current
Stock Ponds	12,500	--
Water Quality	Unknown	Unknown

Methods had to be devised for the capture and processing of both historic and current records considering data quality, economics, time requirements, and including the necessary identification system to provide flexibility in access and retrieval. A more detailed description of the overall data bank system, including procedures, follows in a later section.

A review of the material in table 1 indicates that both numeric and descriptive data exists. The format of the descriptive data for a well is considerably different than that required to describe a dam or a water right. The wide variation in descriptive data required special consideration in selecting record formats to be used in the Colorado Water Data Bank.

The MARS VI DBMS will handle only fixed length records and thus several different subrecord types were defined which allow processing of what might be considered a variable length record. In the case of diversion records, the numeric data for some ditches were recorded daily; however, in other instances the amounts were recorded periodically, or in some cases, lumped as monthly values. Due to the legal requirement that the Data Bank must be able to exactly reproduce the observed historic records, it was essential that a record format be devised which would allow retrieval of actual observed amounts. To satisfy this legal requirement, strict control on data accuracy and number of records in the Data Bank is maintained.

The amount, type and format of water data varies from state to state and thus a standardized water data bank for all states is not feasible. Although the specific elements to be included in a record format may vary, it is felt that the logic and philosophy which are the basis for the CWDBS could be applied to other states.

3. Selection of Data Base Management System

From the outset, it was apparent the Colorado Water Data Bank Project would require a Data Base Management System (DBMS). Because of the time frame specified in the contract, it was not feasible for the Data Bank Project to write its own DBMS and a search of private vendors having available software was undertaken.

The selected DBMS had to be available for the Control Data Corporation (CDC) 6400 computer owned by Colorado State University. This computer system had at that time 65,000 decimal words of central memory; five 841 disk drives with public packs, three of which could be used for permanent file storage; and five 7-track tape drives.

Four candidates for use as a DBMS were found. They were: (1) Remote File Management System (RFMS) from the University of Texas at Austin; (2) SYSTEM 2000, marketed by MRI Systems Corporation of Austin, Texas; (3) MARS VI Version 2.1, marketed by Control Data Corporation; and (4) SISTER, marketed by Temple University. Two of the systems, RFMS and SISTER, were judged to be impractical because of the extensive programming effort required to make them operational. An extensive evaluation of SYSTEM 2000 and MARS VI was carried out by personnel at the Colorado State University Computer Center. The evaluation is described in detail in a project technical report by McMillin [1].

The MARS VI DBMS was chosen over the SYSTEM 2000 DBMS. In general, it was felt that the MARS VI DBMS more closely adhered to industry standards. When the Colorado Water Data Bank Project began operation, on July 1, 1972, the Conference on Data Systems Language (CODASYL) Data Base Task Group (DBTG) "April 1971 Report" was barely a year old. Personnel on the project felt that there was a need for a standardized data base management system. The DBTG Report proposed such a system. While MARS VI certainly did not adhere to the specifications of the report, its file structure was somewhat compatible. Control Data had made a corporate committment to develop and implement a DBMS which was compatible with the DBTG recommendations to CODASYL. This product is known as QUERY/UPDATE.

The MARS VI DBMS has a data base structure which allows user programs to access the data base either through the MARS VI DBMS or by using an entirely external program. This was an important factor in the choice of MARS VI.

4. Characteristics of the MARS VI DBMS

There are several characteristics of the MARS VI DBMS which should be discussed in order that the reader might understand the functioning portion of the Colorado Water Data Bank System (CWDBS). These characteristics have a bearing on the internal structure of data in the Colorado Water Data Bank (CWDB).

1. FILE STRUCTURE - MARS VI has an index sequential file structure with multiple key capability. This results in a partially inverted data base. Those data elements declared as keyed items may be used to make a direct access of all index sequential records containing the keyed value.

2. TABLES - MARS VI maintains a set of internal tables. The internal tables contain unique values for all items which have been declared as keyed. Associated with the unique values are pointers to the index sequential records containing these values.

3. FILE RESIDENCE - The MARS VI DBMS may access data through Rotating Mass Storage (RMS) files or from magnetic tape files. The RMS files may be local non-permanent or permanent files.

4. PROGRAM INTERFACE - A MARS VI data base may be accessed by user programs written in COBOL. The MARS VI DBMS does not communicate directly with these user programs; however, interfacing subroutines are available which enables the data base created by the DBMS to be accessed by user programs written in COBOL.

5. VARIABLE LENGTH RECORDS - The MARS VI DBMS has a limited capability for handling variable length records. Each record type which is of a different length must be on a separate index sequential file. MARS VI allows ten of these files which may be managed concurrently and collectively as a data base.

6. DATA DEFINITION LANGUAGE - MARS VI has a Data Definition Language (DDL) which is used to describe the format of the data elements on each record file. The definition is used by the MARS VI DBMS in all subsequent uses of the RETRIEVAL and UPDATE modules.

7. RETRIEVAL CAPABILITY - Data may be retrieved from a MARS VI data base in two ways. The first method allows the user to retrieve data and process the retrieved data using the MARS VI DBMS directly. This makes use of a RETRIEVAL module followed by a REPORTER module, which allows selected data items to be printed in a very readable format with a minimum of report formatting effort. Basic statistics are also available through the use of these two modules. The second method of access allows the user to retrieve data directly from the data base using the MARS VI DBMS, which writes a sequential file. The sequential file of retrieved data may then be processed by user programs.

8. USER PROGRAM DIRECT ACCESS - Should the user not desire to access the data in the MARS VI data base by using the MARS VI/COBOL interface or using the MARS VI RETRIEVAL module, he may access the index sequential file directly. That is, a user program written in a language such as FORTRAN or COBOL may read the sequential file portion of the index sequential file directly. Thus, when it is desireable, user programs may access data stored in the data base without using the MARS VI DBMS.

5. Development of the Colorado Water Data Bank System (CWDBS)

Project personnel were required by the first year contract to incorporate existing computerized water rights data into the data bank within the first six months. Capture of other historical and current records had to be initiated within the first year. These requirements prohibited initial development of the overall CWDBS and an interim procedure was implemented for storing and capturing data while correction, update and verification procedures were not addressed until the complete system design was initiated in the second year. It was imperative that the project demonstrate its capability by implementing a data base using the MARS VI DBMS.

The water rights data existed on magnetic tape and had been pre-edited and verified and it was possible to directly input these data into the MARS VI DBMS without editing. Updating and correction procedures were tried with this data base and it became apparent that development of the overall CWDBS was imperative to success of the project. Because of personnel limitations, an outside consultant, Fritz & Associates, of Ft. Collins, Colorado, was retained to design a system which could be used for capturing, editing, verifying, updating and retrieving data from the CWDB. The consultant was retained for three months and at the end of that period, submitted a report, Fritz & Associates [2], which was to serve as the working document for further development of the CWDBS.

Implementation of the CWDBS began in July, 1973. Software requirements necessitated some minor modifications to Fritz's system design. Implementation of the system has clarified the user/machine interactions and has allowed development of some universal software and procedures which have been used to process several types of data. This has minimized software overlap and has standardized user procedures for coordinating data capture, correction, verification and updating.

5.1 Structure of Record Formats

Each of the data types listed in table 1 and illustrated in figure 1 has a different length of record to be stored. Because of these variable record lengths, it was decided to implement each of the data types as a separate index sequential file within the MARS VI DBMS. Because of user requirements, it was necessary to be able to cross-reference data between the index sequential files. That is, having used some criteria to select a data record on one index sequential file, it may be necessary to retrieve several associated records from one or more other index sequential files. The MARS VI DBMS allows this type of access to a data base provided a common identifier is specified on each index sequential file in order to link the two types of data.

For example, (reference figure 1), it may be necessary to retrieva all diversion data for a given water right. This data access might be compared to a personnel data base where

the financial records are on one file and address information is on another. The social security number would be the common key to link these files together.

The State of Colorado has developed its own identification system. For administrative purposes, the state is divided into seven large geographic areas where each represents a major river drainage basin. These areas are called divisions (DIV) and each of these is further subdivided into smaller drainage basins called water districts (WD). There are 80 WDs in Colorado. Within each WD a unique five-digit number is assigned to each data collection point. The WD number, when combined with the data point number, creates a unique common identifier (ID) for each data collection point. Using the ID, it is possible to access interrelated data elements from different files in the same retrieval.

Several different types of data may be associated with a single data collection point; e.g., water rights, diversion and water quality. The assignment of the unique ID for the collection point allows the desired cross-referencing and also eliminates the need to assign a different identification number to each record for every data type.

a. Choosing the Keyed Items

The MARS VI DBMS allows a partially inverted file structure. For those data elements within an individual record that the user desires to directly access, MARS VI creates data base keys. The data elements which are chosen as MARS VI keys are said to be inverted and unique valued tables are constructed for each of them. Relative pointers to the index sequential file are constructed for each of the unique values within the corresponding table. Retrieval of data elements which have been inverted requires only that the unique value be looked up in the index tables and the relative position in the index sequential file obtained. The MARS VI DBMS may then directly access the record or records containing the desired value.

For each keyed data element within a data record, on-line storage will be needed for the index tables in addition to that required for the sequential file. MARS VI DBMS users must be careful in the selection of keyed items to provide random retrieval and update capability without increasing the storage requirement excessively.

For the CWDB, three basic data elements were chosen to become keyed data elements in nearly every record type. They are Division (DIV), Water District (WD), and the common identifier (ID).

The primary reason for making a data element a keyed item is to facilitate either updating, retrieval or a combination of both. Within a record there may be data elements that lend themselves to being keyed items for that particular data type; however, these elements may not be common to all record types. To reduce storage and simplify the data base definition, it may be desireable to change some keyed elements to non-keyed elements following the correction, verification and updating of specific data bases. Such a condition is described in section 5.2.

b. Mapping Identifier Numbers

The implementation of data from federal data bases requires that at least a Colorado-assigned ID be inserted into each data record. This is necessary for cross-referencing. There is no standardization between the chosen collection points of the federal data network and the state-chosen collection points for the Colorado water data network. The collection points of the federal network that the State chose to use are a small subset of the entire Colorado data network.

The mapping process whereby a federal ID is mapped to a state-assigned ID to facilitate cross-referencing data within the Colorado Water Data Bank is not a complicated one. However it does seem that this step is unnecessary and would not be required if there was a standardized method for assigning IDs to data gathering networks. Currently, water data captured under federal control may be obtained by all state agencies and cross-referenced through the federal identification system. In some cases, state agencies supply data captured under state control to the federal data base system. In these cases, the state agencies have cooperated and used the federal ID system. What is not easy to do is to make use of federal

data in conjunction with state data. Even more unfeasible is to share data between state agencies. For example, sharing of diversion data between the states of Colorado and Wyoming would be most difficult at this time. Both states have a different identification system and it's not clear whether the respective state agencies address the same type of data as being diversion data. For engineers who have computer modeling applications, it would be most desireable to be able to interchange water data at all governmental levels.

5.2 Working Versus Official Data Base

As indicated earlier in this paper, much of the data in the CWDB which is collected by the state is intended to be a legal record. In order to make this data a legal record, there is an extensive verification process. This process is described in detail later in this paper. To facilitate this extensive verification process, the Colorado Water Data Bank Project has developed the concept of a working data base and an official data base. The working data base contains both verified and non-verified information while the official data base contains only verified records. The structure of these two data bases may differ considerably.

It is intended that the working data base be smaller than the official data base. The working data base contains only that data which has not been verified by the agency or individuals responsible for data capture. Once verified and declared to be correct, data will be transferred to the official data base. A primary difference between the two types of data bases is that the structure of the working data base allows it to serve as both a data base which can be "read from" and a data base which can be "written to". The working data base may be updated by adding new data or by correcting existing data within the data base.

The official data base is thought of as a "read only" data base. It is intended that the official data base will be accessed only to retrieve data for a user. Data which has been verified in the working data base may be transferred and added to the official data base. Once data elements become a part of the official data base, it will be most difficult to make changes to these data elements. Provisions have been made for changes to be made to data in the official data base, but the process involves technicalities much as would be expected in changing any type of legal record. This process is expensive, both in terms of computer cost to perform the updating and time required for an individual to process the change.

In structuring the official data base, several changes have been made in the MARS VI data definition. The changes reflect the fact that the official data base is primarily designed to be read from. Therefore, keys which exist in the working data base for updating purposes are removed. Only items which will be specified frequently for retrieval purposes and those data items that are used for cross-referencing data types are kept as keyed values. Therefore, the storage requirement for the official data base structure versus the working data base structure is significantly less.

5.3 Data Collection Network

The CWDBS identifies three main points in its data collection network. They are: (1) Data collection and verification, (2) The Data Base Administrator (DBA), and (3) The computer software. The data flow between these points is shown in figure 2. This figure details only the data processing for current diversion or current reservoir data. Other types of data employ variations of this data processing procedure.

Data enters the CWDBS from two sources. The largest source is from within Colorado. The second source is from other agencies such as the U.S. Geological Survey. The discussion below presents the collection of data from each source. The acronyms correspond to those used in figure 2.

a. Colorado Water Data

Two points are identified in the network for capturing and verifying data. They are the water commissioners (WC) and the office of the Division of Water Resources (DWR). These two points in the network are primarily responsible for the coding of new data, coding of data

corrections, and verifying data which has been entered into the CWDBS. Modes of data capture include the coding of OpScan mark sense forms and load sheets. Both the WC and DWR must transmit the captured data to the data base administrator (DBA).

The water commissioners are involved in a hierarchical structure. Therefore, the network necessitates their submitting the captured data to the Division Engineer's office, (DIV). Under the control of each of the seven Division Engineers' offices are several water commissioners within the different water districts (WD). Each Division Engineers office is responsible for batching all data submitted by water commissioners in his division. The data is transmitted periodically to the data base administrator (DBA). Water commissioners capture only current diversion or reservoir records.

Historical data is captured by the State Engineer's office, DWR, and is batched and transmitted directly to the DBA. This data is also captured utilizing either the OpScan mark sense sheets or load sheets.

The data base administrator (DBA) is responsible for logging and submitting data received from either DIV or DWR. This data is received in either OpScan or load sheet form. The OpScan data is submitted to be captured on the OpScan 100DM to 7-track tape. Load sheets are submitted for keypunching. The DBA is then further responsible for maintenance and updating of the CWDB. This is accomplished by using the CWDBS computer software.

After the data base has been updated, the DBA is responsible for distributing either error lists or the verification reports published by the CWDBS software. This distribution process involves returning the reports and error lists to the respective point in the data network from which the data originated. Therefore, these reports are returned either to DIV or to DWR. If the report and error list are returned to DIV, they are then further distributed to each WC. In the case of DWR, which is an originating source, no further distribution is required.

At each originating DWR or WC, additional manual processing is performed. In the case of edit error lists, each error is resolved. Corrections for the errors are coded and the processing begins a new loop.

In the case of the verification reports (see fig. 4), the originating source must check the data values associated with each data element in the report. The report is verified on a page-by-page basis. On each page is a signature block (no. 9, fig. 4), which is signed by the individual who coded the record for original input. The signature is affixed to the verification report page only if all data on that page is correct. The data on that page is then eligible to be moved to the official data base and the report is forwarded to the DBA. Should there be errors on the page, then corrections must be coded for the incorrect data. These corrections then enter the data processing loop.

It is up to the DBA to determine when a logical batch of data from the working data base has been verified as being correct. At the discretion of the DBA, the data from the working data base is moved to the official data base. At the same time, the signed verification reports are distributed to DWR to be entered into the archives as an official legal record. The CWDBS software is responsible for removing the data from the working data base to the official data base.

The object of the verification process is to move all data from the working data base to the official data base. Since data is constantly being captured, this objective seemingly may never be reached. However, the capturing of water data in Colorado is oriented around an irrigation year which begins on November 1 of the first calendar year and continues through October 31 of the following year. Therefore, it is intended that on or about October 31 the old working data base will be "frozen" and a new working data base will be initiated. It may take a few weeks into the new irrigation year to remove all remaining data from the previous year's working data base.

b. External Water Data

Not all data entered into the CWDB is data which has been captured under state control. Data may come from separate state agencies or a federal data collection agency. When entering this data into the CWDB there may or may not be a verification process. For the most part, this data is accepted at face value. However, general editing for obvious data errors is performed in the data processing system.

In lieu of the working data base concept, which is required for state gathered data, there are intermediate data files generated for external sources of data. Generally, this intermediate data file represents the procedure of extracting only the needed data from the external source and mapping the state assigned identifier to the external data. In some cases, data conversion or modification may take place. The resulting external data file then is loaded directly to the official data base. In keeping with the concept of the official data base, it is not intended that external data appearing in the data base will be modified. Data may be added through an add-on load.

5.4 Computer Software for the CWDBS

The software which the DBA uses to maintain the CWDB is written in two computer languages in conjunction with the MARS VI DBMS. Programs exist in FORTRAN and COBOL as well as input specifications to the MARS VI DBMS. The CWDBS software obtains most of its control information through user-supplied tables. These tables are maintained by the system by entering table information as data. Header information identifies the data as tables and the tables are updated.

The use of tables allows the user more control over the CWDBS software. Old record formats may be changed and new record formats added without software modification.

Figure 3 presents the general flow of data through the CWDBS software. The acronyms presented here correspond with those in the system flow diagram,

a. DBAC--Preprocessor

Program DBAC is responsible for processing the data to be input into CWDBS. This initial processing involves reading of data from external magnetic tape sources, 80-column data cards, or magnetic tape generated by the OpScan 100DM. DBAC reads the data from these sources and adds unique sequencing information to each record from the input source. Header records precede each type of data to be entered into the system. Information on these header records, combined with a sequential numbering system, creates a unique identifier for each data record.

In the case of OpScan input, a further requirement for DBAC is that it unscrambles and decodes the magnetic tape input which is generated by the OpScan processor. The UNSCRAMBLE/DECODE software is table-driven and these tables exist on a permanent file accessible by program DBAC.

Program DBAC then sorts the output by data type and generates a 7-track magnetic tape of this data. A disk file of the data is used as input to a subsequent program DBAD in the CWDBS.

b. DBAD--Edit/Update

Program DBAD is responsible for all data editing. This editing is done within the data record at the data element level and between data elements. As records are edited, they are either accepted or rejected. The rejected records are written to an edit reject file which exists on 7-track magnetic tape. The accepted records are processed.

Processing of accepted records involves direct updating of the CWDB through the MARS/COBOL interface software or indirect updating through the MARS VI DBMS. Direct updating can only be done to those data elements which are non-keyed. If the non-keyed data being

entered into the system is original and no record exists where this data can be added, a record is written to a MARS VI transaction file. The transaction file will be processed by the MARS VI DBMS. If updating is to be done to keyed data elements, then program DBAD generates a MARS VI transaction file which will later be input to the MARS VI DBMS. Program DBAD is also responsible for updating the control tables which are used by all CWDBS software. Updating of these tables makes use of random access files. DBAD does not use the MARS VI DBMS to update the tables.

As part of the updating process, program DBAD compares the update transactions against the edit reject file. The software is capable of performing modification to the edit reject file to correct the errors that occurred in the data records when they were written to the edit reject file. When a data record on the edit reject file is corrected, it is removed from that file and input into the normal data base edit and update procedures. The objective is to eventually remove all records from the edit reject file.

In addition, program DBAD is responsible for writing the edit error report. These error lists are taken by the DBA and distributed to the peoper points in the data collection network.

c. MARS VI--Update/Add-On Load

The MARS VI DBMS is utilized to update the data base for keyed data elements and new data records. Updating of keyed data elements is necessarily more expensive and experience has indicated that updates to keyed data elements should be batched together. This is because it is less expensive to update five keys in one session than to do so in five sessions.

The new data records are processed through the add-on load feature of the MARS VI DBMS. This is the most common type of update transaction.

d. Report Generation

After the updates have been processed, the MARS VI software reads retrieval specifications from a card data file. Retrieval from the CWDBS may be done using either the MARS VI DBMS or the MARS/COBOL interface. MARS VI provides the user the capability of having a quick look at data in the data base. Using the MARS VI RETRIEVAL and REPORTER modules, the user can create reports in a short time. However, because of format and logic limitations of the REPORTER module, most of the project's reports are created using special report generation software.

A sample report generated by the special report software is included as figure 4. This report is a complex report requiring cross-referencing of multiple index sequential files, data computations, and data interpretation. Item 1 indicates this report is for an irrigation year. Items 2 and 4 specify location. They require three accesses of two files. The structure number (03551) in item 2 requires an access to a location file to retrieve the structure name. When retrieving the stream number (001) in item 4, an access is required to the location file to obtain the stream name. The information obtained in item 3 requires yet another access to a file. The names in item 5 are stored in tables within the program. Item 6 indicates observed data as indicated by the asterisk. Observed data is the only data entered into the CWDBS. However, the report requirements state that if data is missing then values are to be interpolated from the last observed value. That is, the last observed value is carried forward until the next observed value. This is indicated by item 7. Item 8 indicates the computations which are performed. If a verification page is correct, the page is signed in the lower right corner as indicated by item 9.

The report generation software is responsible for generating all verification reports for the CWDBS. This software is required to produce quite complex reports. There is often a requirement to merge data from several of the MARS VI index sequential files. The MARS VI DBMS is used to retrieve the desired data and the report software reads the intermediate files to produce the reports.

5.5 Verification Procedure

The verification software of the CWDBS is imbedded almost entirely in program DBAD. Since it is a stringent requirement that the data captured under state control be verified as being absolutely correct and entered as a matter of legal record into the archives, this software is quite important. Basically, the software must keep track of the status of each data element in the working data base. The possible status conditions are: (1) The data element has been entered into the data base but no verification report has been produced for it, (2) The data element has been included in a verification report and is assumed to be correct, or (3) The data element has been corrected through use of a verification report.

The general logic is that a data element enters the CWDB as status 1. When this data is included in a verification report its status is changed to status 2. Data elements which have a status of 2 are assumed to be correct. Should the verification report reveal that a data element is in error it is corrected through the CWDBS software. At this time its status is changed to status 3. Additional verification reports are produced on status 3 data and the status is changed to 2 again. The data element is again assumed to be correct, until reported to be in error. The goal is to have all data with a status 2. The DBA will determine after receiving signed verification reports when to move status 2 data to the official data base.

An aesthetic problem exists in having an individual sign the verification report and its becoming a legal record. What guarantee does this individual have that the data he verified as being correct was the actual data (combination of binary zeros and ones at the most elementary computer level) that was transmitted to the official data base? It has been suggested that in order for the data to become an official record, that an additional report must be produced from the official data base after the data has been moved from the working data base. Currently, these problems are still being resolved between data bank personnel and DWR. Basically, the problem is to what degree can one trust computer software? If the computer software is 100% logically correct, to what degree can computer hardware be trusted?

6. Conclusions

The Colorado Water Data Bank Project has been in operation approximately 1 1/2 years. The major effort during this period has been to develop the logic, procedures, and programs to be incorporated in the overall Colorado Water Data Bank System (CWDBS). The Control Data Corporation MARS VI Data Base Management System has been incorporated as an integral part of the overall system. Several different types of records have been captured and placed in the data bank and more recently requests for access and retrieval of data have been processed. As would be expected, the project has experienced both success and setbacks on meeting certain objectives within the selected time frame.

Although complete implementation of the Colorado Water Data Bank System is not expected prior to June, 1975, considerable progress has been made and it is possible to draw the following conclusions:

1. There is a need to have localized water data banks which will contain many different types of data and which will provide the capability for access and retrieval of all the information required to make administrative or management decisions at one time. Centralizing the data location will minimize retrieval costs, allow cross-referencing, and provide the information within a reasonable time frame.

2. Currently, it is not possible to directly interchange water data with other federal or state agencies. While variations may be subtle, each agency has implemented its own identification system for the data collection points within each agency. Data from other agencies may be entered into the CWDBS. However, the external agency identifier must be mapped to a Colorado identifier.

3. It is possible to utilize a commerically available data base management system as an integral part of a complex water data bank system. Utilization of a commercial data base management system requires standardization of input and output procedures. Utilization of the MARS VI DBMS permitted this project to begin capturing of data at least six months earlier than would have been possible if all programs had been written by project personnel.

4. The wide variety of data to be placed in the data bank has required the establishment of several record files with different record formats. Organization of these files has permitted the treatment of both fixed and variable length records.

5. Although data was originally captured using interim procedures, development of Colorado Water Data Bank System has provided the software and procedures for capturing different types of data with a minimum of effort by the Data Base Administrator (DBA).

6. Data is being captured with mark sense forms, allowing the water commissioners to prepare the machine readable document, thus minimizing transferral errors and time required to put the data in the data base.

7. The processing of both water rights and diversion records has used the overall CWDBS. During the next year, the necessary tables and edit routines will be defined and incorporated into the system for the other types of water data. Development of detailed user documentation is underway which should allow the Data Base Administrator (DBA) to process all incoming data as well as honor data retrievals.

7. Acknowledgements

The authors wish to acknowledge the funding of the Colorado Water Data Bank Project by the State of Colorado. The Colorado Water Data Bank System described in the paper was developed jointly by personnel from Colorado State University and the Division of Water Resources. Correspondence on this paper may be addressed directly to the authors at Colorado State University; however, programs, data, or report requests should be sent to the Division of Water Resources, 1845 Sherman Street, Denver, Colorado, ATTENTION: Chief, Computer Services Branch.

8. References

[1] McMillin, Norval E. Evaluation of a Data Base Management System for the Colorado Water Data Bank Project. Colorado Water Data Bank Project, Colorado State University, Technical Report Number 72-01, October 6, 1972.

[2] Fritz, Terrence L. Proposed System Design of Manual and Automatic Data Processing Procedures for Colorado Water Resources Data Bank Project. T. L. Fritz and Associates, Fort Collins, Colorado, June 26, 1973.

9. Addendum

Some modification of the preliminary draft was made for clarity and to emphasize discussion points raised during the symposium. Questions forwarded to the authors are presented and answered in this addendum.

Question: Why can you not interchange data with Wyoming or Nebraska:

Answer: Assuming data exists, it can be exchanged; however, the data format would probably not be compatible between states. The record formats for two different states might contain different data elements. Most likely location identifier for the data collection points would reflect individual state location systems and would not allow physical referencing of a point in one state to a similar point across

the state line in another state. In some instances reformatting of data records to a common format will allow compatible usage of the information, but the lack of uniformity in the data elements included in each record can not be easily overcome. See section 5.1 and 6.0 for more discussion.

Question: Could you elaborate on the conflict of federal data and state data?
Answer: The federal versus state data uniformity problem is similar to that discussed above between two states. Also see section 5.1.

Question: What is the function of the Data Base Administrator (DBA)?
Answer: Currently, one individual performs the function of the DBA. All transactions which will update or modify the data base must be processed by the DBA. In addition, requests for verification reports must be submitted to the DBA for processing.

Question: What is the error rate using the OPSCAN mark sense technique? What was the degree of acceptability by the users of the mark sense forms.
Answer? We have found the existing OPSCAN machine to adequately capture the marked forms with a very low machine reject rate, a small fraction of one percent. Existing edit programs and the verification procedure define miss-marked or improper data. Numbers to evaluate the errors due to improper marking versus machine read problems are not availabe, but our success in data capture has encourage us to use the same mechanism for another year. Some attempt will be made in June, 1974 to evaluate the relative merits including cost, of mark sense capture versus key-punching.

Education programs were held to acquaint personnel with the mark sense technique. Assuming data codes and marking procedures are well defined prior to education meetings, most of the personnel have adapted to mark sense capture of data. Some redesign of forms has been undertaken to incorporate suggestions from the users. Careful layout of forms is most important to the success of the technique.

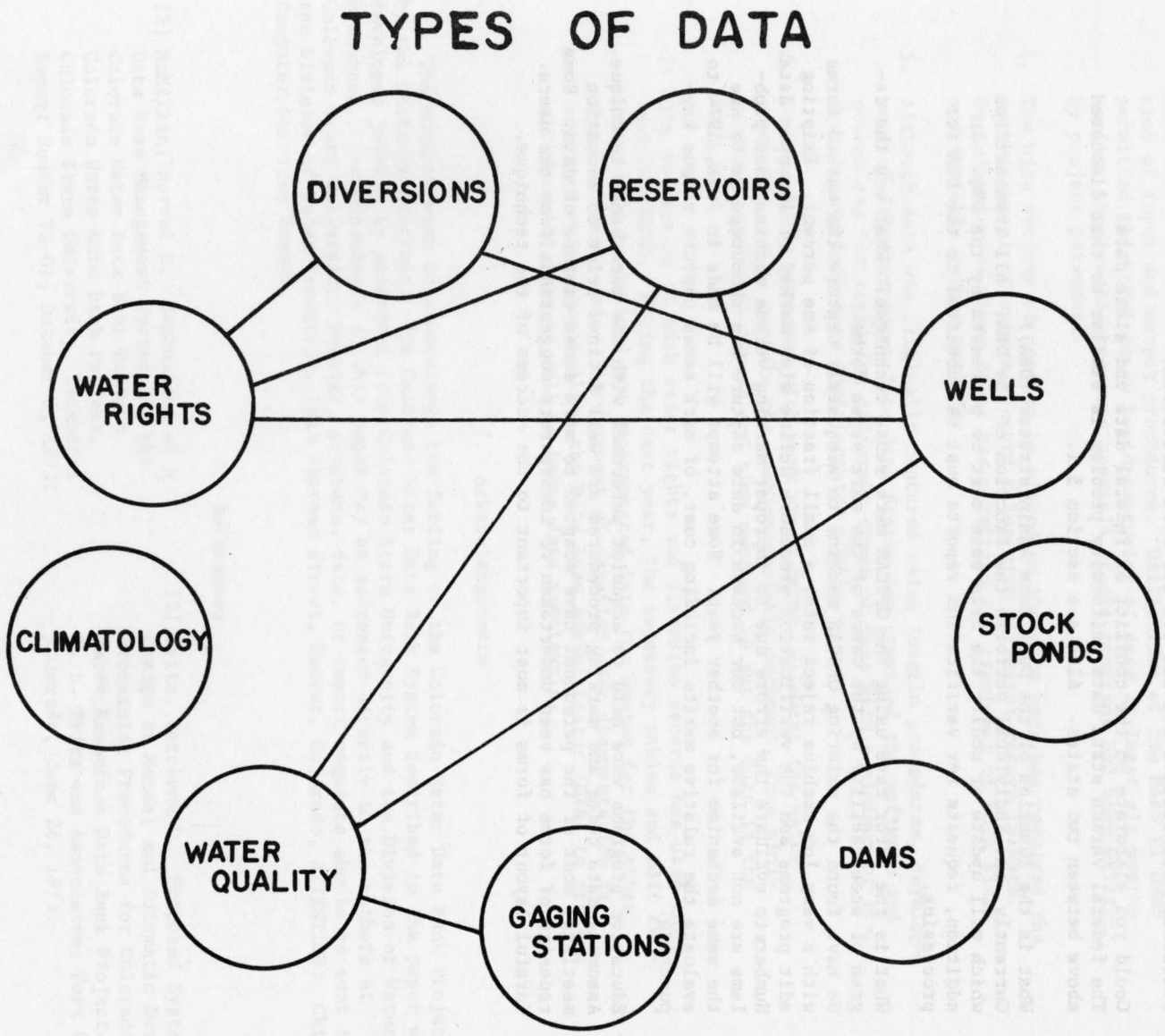

Figure 1. Types of data to be included in the Colorado Data Bank showing those types to be cross-referenced.

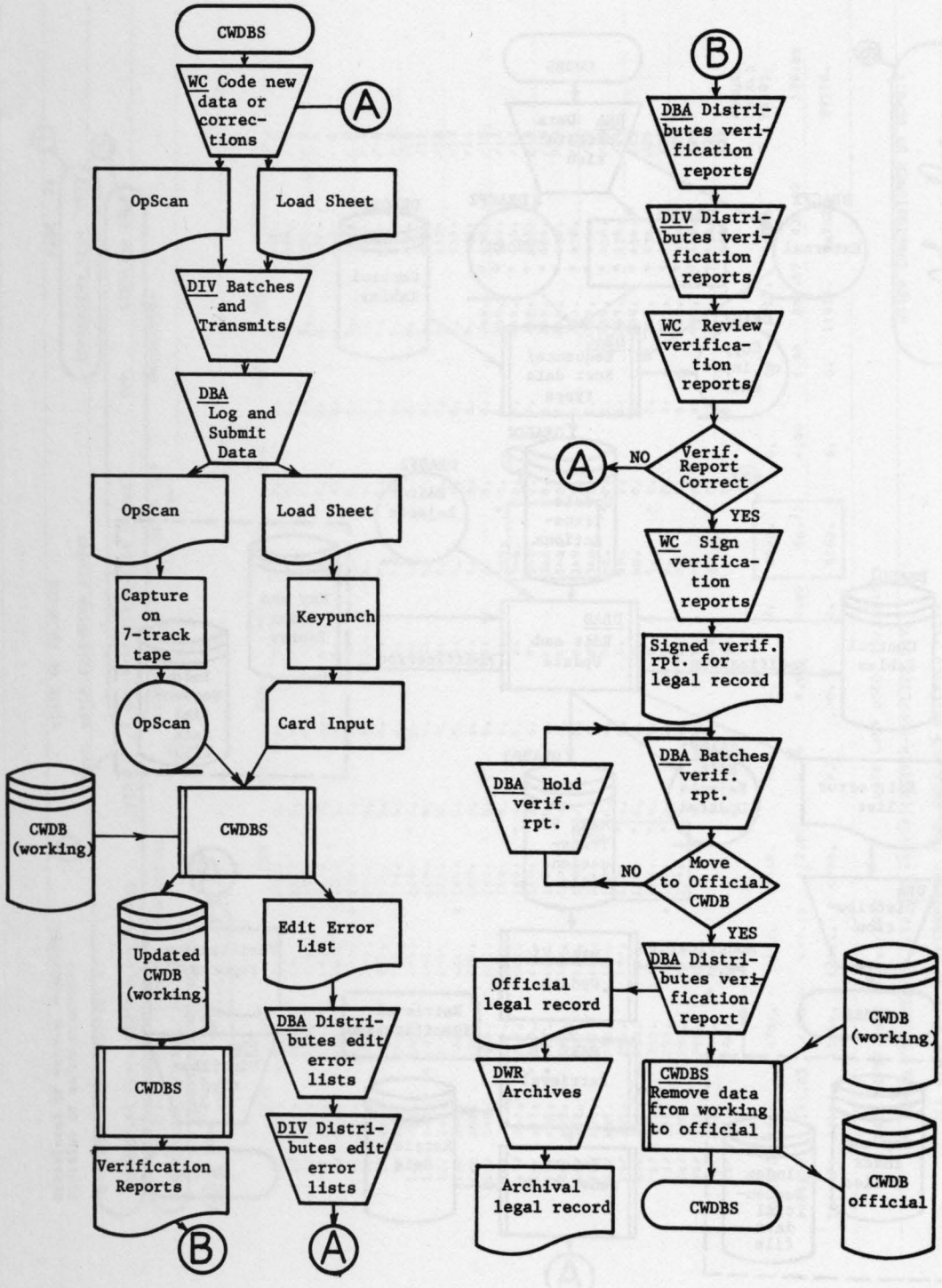

Figure 2. Flow diagram illustrating procedure for capturing current reservoir or diversion data.

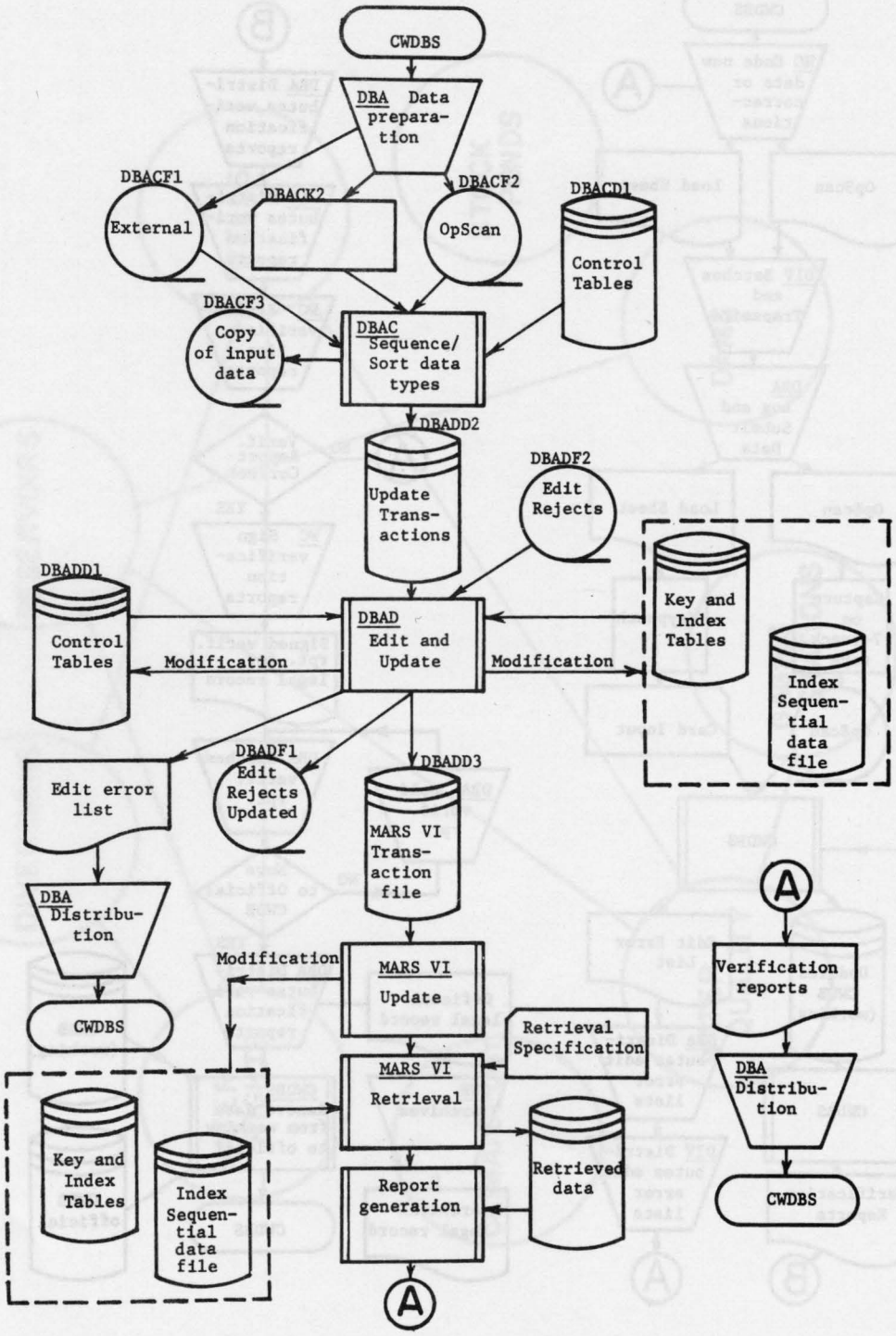

Figure 3. Skematic showing connection between software programs.

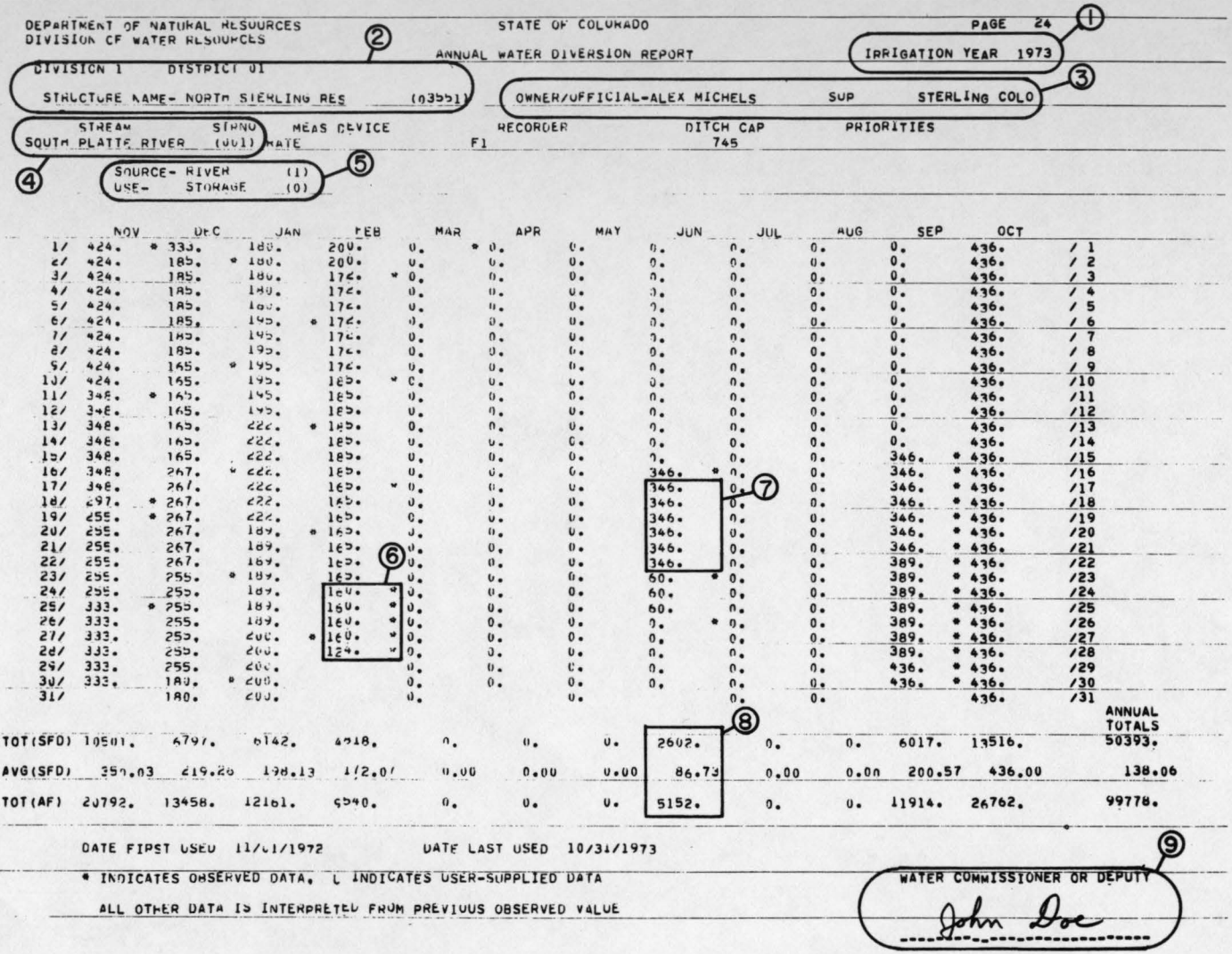

Figure 4. Typical annual summary and verification report for diversion records.