Ground Water Data : Storage & Retrieval System

1. Introduction

The increasing capability and availability of computer-related technology are having a revolutionary effect upon the techniques available to those responsible for the assessment and management of natural resources.

The most fundamental advantages of computerized processing systems, and perhaps still the most significant, are:

- Data are stored in a compact, organized manner.
- Data may be rapidly retrieved in a wide variety of formats and aggregations.

With respect to subsequent data applications there are other important benefits. The computer's ability to perform systematic consistency checks offers a powerful aid to improving and maintaining the quality of data and data analysis techniques may be much more sophisticated than those found in any routine manual processing system.

The many potential benefits of computerized data processing may only be fully realized by proper planning and management. This involves equipment (hardware), computer programs (software) and personnel. Unfortunately, there are still problems of compatibility between different computers, and this incompatibility affects both hardware and software. Care needs to be taken at all stages of system development in order that such problems may be minimized.

2.0 Data Operation

The first group of operations is associated with data collection, and covers the stages between data observation and its arrival at some processing centre.

The second group of operations relates to data processing and covers the preparation, input and quality control of the data, the updating of the data base (or data bank), techniques to enhance the utility of the data, and standard data retrieval options.

Lecture delivered by Sri S.K. Singh, Scientist 'C', NIH,

The final group of operations are those associated with data application, and involve the use of data in subject specific analyses such as the estimation of aquifer recharge, land use evaluation and efficiency of fertilizer treatments.

The scope of this chapter is limited to the operation of second group only. A brief idea about the various operations of this group of data processing is given here under.

2.1 Data processing

It is the operations to be performed in data processing which are the focal point of this chapter. The data-collection exercise provides a variety of text descriptions, numerical values, charts, computer compatible media, etc. It is the task of the data-processing system to convert these diverse data, recorded on diverse media, into integrated data sets structured to suit the information requirements of the data user. Table 1 shows the main components of a data-processing system.

2.1.1 Data preparation

This activity comprises the operations necessary to convert data from the format in which it is received to a format suitable for input to the computer. The complexity of these operations is obviously governed by the degree of computer compatibility of the recording media. Much handwritten data may need to be transcribed and perhaps replaced by some coded value. This produces punching documents (sometimes called coding sheets) from which the data are subsequently key punched directly onto the computer or onto some computer compatible storage media, e.g. punched cards, diskettes or magnetic tape.

This preparation is both tedious and liable to produce transcription or coding errors. There are methods for reducing or eliminating data preparation effort. These methods include the design of field data sheets from which keying may be performed directly.

2.1.2 Data entry

There has recently been a major shift in methods of data entry (input). Punched cards and paper tape have been widely replaced by key-to-tape or key-to-diskette systems. Large-scale keying operations normally comprise type stages; in the first (punching) stage data is keyed by one operator, in the second (verifying) stage the same data is keyed by another operator and compared with the original input. Any differences between the two data sets cause the keyboard to look, and the data value being entered can be checked. This system is very successful in eliminating data entry errors.

There are other more specialized data entry methods, e.g. the conversion of charts and maps to digital format using manual or automatic "digitizers".

2.1.3 Data validation

Once entered into the computer, data should be subjected to a set of checks designed to identify incorrect or unusual data values. These may be simple checks that compare the value entered with an expected range for that parameter, or more complex ones that compare, say, daily rainfall totals at one site with totals recorded at adjacent sites. If data has been coded, code values and code combinations can be checked for validity.

Mis-coded or mis-punched data identified in this way can normally be easily corrected. However, correction of suspect data values, or the querying of missing values need to be referred to those responsible for data collection, an important reason for allowing this initial phase of data processing to be carried out by the field observers themselves.

2.1.4 Primary processing

This stage of processing is concerned with preparing and assembling the data in the format necessary for it to be added to the existing data base. This may include standardization of measurement units, additional levels of coding for storage purposes, and the estimation of derived parameters, e.g. estimation of Penman evaporation from climatological data. The amount of primary processing necessary is related to the degree of coding of the data before input. Systems which utilize less initial coding are easier for the extent to which manual coding is used is governed by the quality and availability of data preparation and computer programming staff, and the capabilities of the computer system used.

2.1.5 Data base updating

Having processed the data to the correct format, it may be incorporated into the data base using an updating program. It is usual to update the data base at some fixed time interval, monthly being a typical interval for natural resource planning data. This requires the allocation of space on the various storage media to hold recently input data until the next update run.

The data base comprises sets of data cross-references in a way which reflects the relationships between various data items and data groups. This structuring of data sets is fundamental to the proper design of any data base.

TABLE 1

The components of data processing

D A T A P R O C E S S I N G							
DATA PREPARATION	DATA ENTRY (INPUT)	VALIDATION	PRIMARY PROCESSING	DATA BASE UPDATING	SECONDARY PROCESSING	RETRIEVAL	OUTPUT
Prepare punching documents by: 1. Transcription Field note book entries Non-standard data formats 2. Coding Reduction/ standardisation of input data	1. Punching document a. Direct keying through VDU b. Keying onto computer compatible media Punched cards Key-to-tape Key-to-disk 2. Charts and maps Direct input by digitiser 3. Computer compatible media a. Tapes/cassette b. Diskettes c. Cards d. Communication lines (telemetred data) e. Mark sense/ optical character readers	1. Range checks 2. Sum checks 3. Inter-station consistency checks	1. Standardisation of units 2. Calculation of derived parameters 3. Further coding of input to reduce storage requirements 4. Arranging data in data base format	1. Add new data sets to existing data base 2. Report any errors	1. Programs for routine reports 2. Statistical summaries 3. Infilling missing data values 4. Interpolation or aggregation of data	1. Selection of data by: a. Parameter type b. Parameter value c. Location d. Period of record e. Time interval of record 2. Selection of output device	1. Printers 2. Plotters 3. VDU 4. Computer storage media 5. Microfilm 6. Telemetry
	E R R O R C O R R E C T I O N						

Table 1 shows, as is traditional, that data base updating does not occur until the 5th of eight stages. With this arrangement, validation is done on every "transaction" before it is added to the data base. A recent development associated with widespread interactive processing facilities is for data base updating to be the first stage as part of data entry. Subsequent stages then operate on the data base using its facilities for manipulation and maintenance of data integrity. With this arrangement both management and users may apply the same validation tests, which can be helpful when users' demands for urgency exceed management resources for validation. This use of full data base facilities in regional, even field offices has become feasible with microcomputer software structured identically to main frame software which supports the national data base, although having capacity for less data.

2.1.6 Secondary processing

This stage of processing covers the reports generated and analyses performed on a routine basis after the data base has been updated.

For example it is usual to provide statistics comparing the most recently input values of a parameter with its long-term variation, in order that any trend in values can be monitored.

There are also methods of aggregating or interpolating data, including the infilling of missing data values, which generally enhance the subsequent usefulness of the data, and programs to perform these analyses are normally run at this stage.

2.1.7 Retrieval

An important objective of an automated data-processing system is to provide a rapid and comprehensive response to *ad hoc* requests for data retrieval and interpretation. Thus, a wide range of retrieval options should be available to allow the interrogation of the data base and the selection of ranges and combinations of data, e.g. the selection of all values of parameter H above a value V which were recorded at location L between dates A and B.

Retrieval options should also include the facility to combine selected data sets and prepare data files for subsequent input to standard statistical software or user application programs.

Another attribute of retrieval options is the ability to allocate any suitable computer output device for the presentation of the selected data.

2.1.8 Output

Modern computers support a wide range of output devices. These include, in addition to the range of computer storage media, printers, plotters, and cathode ray tubes or visual display units (CRTs or VDUs - also

identical to a conventional television screen). Of increasing use are devices supporting both graphics and colour options.

The above data-processing operations have been described only at an introductory level and are considered, as indicated, in the folowing chapter of this publication.

As already stated, it may be desirable to split the processing operations between physical locations. It is proposed that initial data preparation and validation are best performed by field staff as a continuation of the data-collection exercise. The use of microcomputers should be considered to assist in this function. However, the computer and manpower requirements for the operation of an integrated land and water data base imply the existence of a central computer facility.

Initial data preparation and validation by field staff using microcomputers as a continuation of data gathering has been implemented in some countries. In such cases, a central computer facility may not be necessary when the advantages of integration are achieved through use of a common data structure to allow free interchange of data between computers. It has recently become feasible to implement on microcomputers a relational data base which supports all the functions provided on the largest computers for resource data - although for smaller quantities.

3.0 The Nature of Data

There are certain universal characteristics of data, which are independent of the quantity or element measured, but strongly influence data collection and storage. Two important characteristics of data are the distribution of the measurements in space and time. In land and water systems there are variations in the values of physical parameters in both space and time.

When monitoring natural resource data, measurements of quantities of interest have to be made at a number of sampling points and when time variation is also involved a series of observations at different times are also needed. Usually to economize on data-collection costs the observations have to be spaced as widely as possible consistent with being able to interpolate intermediate values to an acceptable accuracy. Three dimensions

may be recognized: the place of the observation, the time it was made, and actual parameter(s) observed. Figure 1 shows this three-dimensional nature of data. The entries in the matrix may be the numerical value of some measurement, a quality or property of the place, or some text note or comment.

Thus with every observation of a parameter are associated other information such as the time and place of measurement and the name of the parameter observed which serve to identify the data. A data-storage system has to reflect this identification information in some manner so that the data stored may be used correctly. In many cases the place of storage of the data can provide some of this identification, for example all river stage observations at a particular station may be held in a single file, and the times of observations will be given by their ordering within the file. In other cases the place or time of observation will have to be explicitly stated, for example a bore-hole log record would contain the coordinates of the bore-hole. By grouping the data the amount of extra information which needs to be stored can be reduced. This chapter discusses how data can be broken down into groups for storage, and defines those groups and identifies the attributes of each which govern data storage. Also described are the general characteristics of data-storage systems and the advantages and methods of coding data.

3.1 Data types

To assist in identifying data types, an immediate distinction may be drawn between data which is fixed and data which varies in either time or space. In the case of fixed data only a single entry need be made along the axis of the dimension in which the parameter is fixed. Thus, data which is related to a physical object at a specific location, or to a parameter which is constant in space need to be represented by only one entry on the space axis. Similarly parameters which are fixed in time need only one entry on the time axis. Conversely, variable parameters must be represented by a series of entries along the axis of the varying dimension.

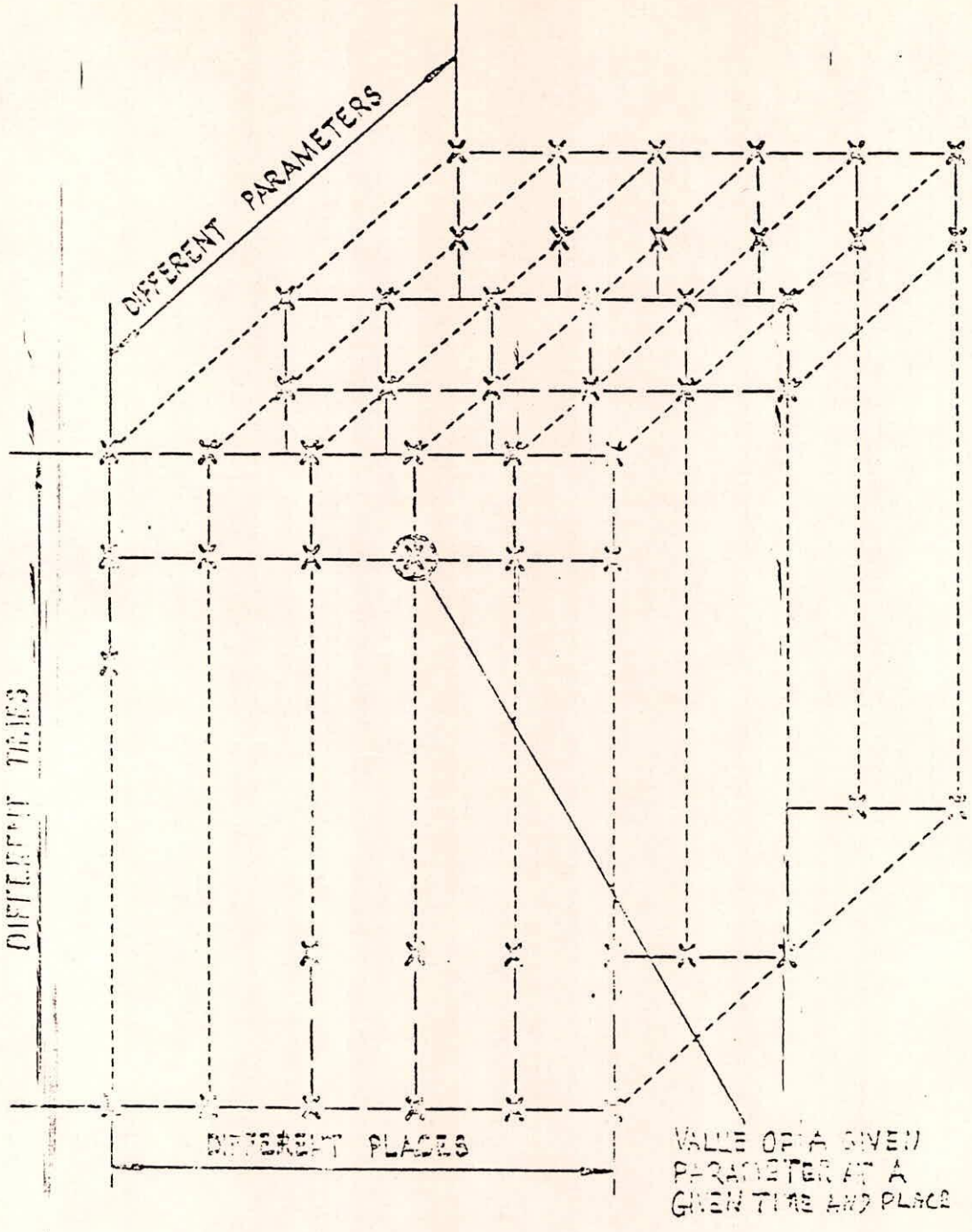

Figure 1 - The three-dimensional nature of resources data

In this context fixed does not necessarily mean absolutely invariant, but varying within a range which, for the purposes of data application, may be considered fixed. For example, the physical, chemical and mineralogical properties of soil are constantly changing, but the rate of change is generally so slow that soil survey data are usually considered as fixed in time. Figure 3 presents the possible combinations of fixed and variable data in the space and time dimensions and gives examples of each type in terms of land and water data. A set of data collected at a point for a parameter which varies with time is called a time series.

Time series data may be divided into three groups which relate to the intervals at which observations are made. There are regular time interval, irregular time interval, and continuous observations. The distinction is drawn as it affects storage formats. Spatially varied data may be divided into two groups to highlight the difference between data which are sampled at points and requires spatial interpolation, and data that are sampled on an areal basis by remote sensing techniques.

Remote sensed data in the form of photographs need to be manually or automatically digitized for data entry. If the data are satellite based they are probably provided directly in computer compatible form. However, such data usually need compressing before they can be included in the normal scale of data base. Similar problems arise with data sampled continuously with time as these also need to be digitized. However, the digitization process reduces the number of data storage formats required, areal data being reduced to sets of x-y distance co-ordinates, and continuous time data being digitized into an interval time series. Thus, referring to Figure 2, computer storage formats are only required for non-time series data types 1 and 2, and time series data types 4, 5, 7, 8.

3.1.1 Time series data

Three kinds of time series have been defined. The regular time interval series, the irregular time interval series and the continuous time series. These series usually relate to a single parameter. However, when several parameters are measured simultaneously at the same site they are additionally described as being a multiple element or parameter series. Examples of this type of series are data from meteorological stations where several parameters are all measured at 09.00 GMT, or the values of water quality parameters derived from a sample taken at a river quality monitoring station.

For data observed at regular time intervals, only the time of observation of the first item is needed when storing the data. The times of occurrence of subsequent observations can be estimated by knowing the standard time interval.

For data observed at irregular intervals, data must be stored as pairs of values. The first value in each pair is a time and the second the observed value at that time. The time may be the actual time of occurrence, it may be the interval between the previous observation or it may be the time measured from some arbitrary time datum. Much of this type of data will be obtained by digitizing charts which converts the continuous time series into an irregular interval series, a stream of pairs of times and values. The time value is given in interval form but may be converted to actual time for storage.

TIME			SPACE		
			FIXED PARAMETER	VARIABLE PARAMETER	
				POINT SAMPLING	AREA SAMPLING+
	FIXED PARAMETER (Non-Time Series Data)		1. O-O-N	2. O-R-S	3. O-C-D
			Physical works: Dams, pumps, wells, boreholes etc.	Soil profile data.	Topographical data from aerial surveys.
	VARIABLE PARAMETER (Time Series Data)	REGULAR SAMPLING	4. R-O-T	5. R-R-TS	6. R-C-D
			Monthly reservoir levels.	Daily rainfall data.	Land use by orbiting satellite.
		IRREGULAR SAMPLING	7. R-O-T	8. R-R-TS	9. R-C-D
			Gauging of ephemeral springs.	River flood gauging.	Land use by aerial photography.
		CONTINUOUS SAMPLING	10. C-O-D	11. C-R-D	12. C-C-D
			Chart recording of reservoir discharge.	Chart recording of borehole water level.	Meteorological data from geostationary satellite.

Notes: 1. Sampling Code: X-Y-Z where X and Y are space and time sampling frequencies, Z is a data interpretation code.
Codes for X and Y values: O – Sampled once only. R – Sampled repeatedly. C – Sampled continuously.
Codes for Z values: T – Needs interpolation in time. S – Needs interpolation in space.
N – Needs no interpolation. D – Needs digitising.

2. + Satellite data may be in digital form but will need compressing.

Figure 2 – Space-time combinations of data

Data in a multiple parameter time series can be stored as a number of separate single parameter series. However, this means duplicating station and time details which are common to all parameters, and in general such data is more efficiently stored as a multiple parameter series. The storage format will be similar to the single parameter series except that the time value will be the sampling time for all parameters that follow, and a code will indicate which parameter values are stored and in which order they appear.

An important basis for classifying time series is by the interpolation method used when the data are stored or drawn. Associated a label which states what KIND of data it is and therefore what type of interpolation is appropriate to estimate values between filed values. In addition, it is useful to define data as INSTANTANEOUS, AVERAGE, ACCUMULATED and INCREMENTAL.

A recorder usually makes a series of measurements, e.g. stage height at various INSTANTS of time, perhaps once every 15 minutes. Such a series is interpolated by sloping lines. Alternatively, a record may contain a constant value over each period of time, e.g. spillway gate openings, which are interpolated as a HISTOGRAM. Mean averages are computed and stored they are of this kind. A recorder may measure the increment over each period of time, e.g. rainfall. This kind of data is described as INCREMENTAL data and retrieval gives the total or the average rate over the period of time that the recorded value was accumulated, depending whether the values are tabulated or plotted respectively. Finally, a series of SAMPLES is a set of simultaneous measurements repeated at such infrequent intervals that there is no valid interpolation for times in between. SAMPLES define calibration functions (RATINGS) that enable transformations, such as stage height to discharge.

3.1.2 Non-time series data

Non-time series data vary only along the space and parameter axes of the basic data set in Figure 1, and are represented by data types 1 to 3 in Figure 2. It refers to points in space, identifiable areas of space, individuals and individual holdings. The parameter values at these points or areas are considered as constant with time and are immensely varied including details of soil, surface and ground water resources, land ownership, and socio-economic data.

This wide range of data types means that, unlike time series data, it is not possible to generalize storage formats, and each case will need to be assessed individually.

There is one kind of non-time series data which requires special treatment. This is typified by soil data and geological log data for bore-holes. In this type of data there is a general description of the location, followed by the description of several soil horizons or geological strata at a variable number of levels. These descriptions may include a variety of chemical and physical analyses of samples taken at each level. A potentially very large amount of data may be described, which makes its storage as a simple array of locations and parameters very inefficient, unless some

data compression is performed. In such a case a hierarchical storage system is desirable where locational details could be stored at the highest level with cross references to separate sub-groups containing data from each sampling depth.

3.1.3 Spatially varied data

Spatially varied data are conventionally presented as maps showing the extent or variation of different parameters. For computer storage purposes the map may be converted into digital form, aggregated into table form, or represented analytically in the form of a mathematical model.

In the case of map data the coordinates of the boundaries of each discretely identified area parcel can be stored. The same method is also used to represent linear features, data usually being abstracted by means of a digitizer. Lines are traced by the digitizer cursor noting the coordinates of the lines at some pre-set frequency. The sampling frequency may be manually or automatically controlled. A digitized map may be replotted on an x-y plotter to any required scale, and a great variety of processing can be carried out such as overlaying different parameters on one another in the same way as overlaying transparent map sheets or identifying intersections of areas with different combinations of parameters. Most software is also able to automatically calculate areas. The main characteristics of this system of storage and manipulating data is its relatively high cost and requirement for special machines. Its advantage is that there is no loss of spatial information between the original and the computer based information.

In the case of tabular data each row in the table may typically represent an area parcel that can be conveniently located on a map. Each column may represent some property of that labelled parcel, such as population density, or the percentage of the parcel devoted to some particular land use. Except in certain special cases of the table the form and spatial relationships of the parcels of land are lost. This is certainly true at the simplest level when no locational information is given. At the next level the locational reference may be given as the approximate centre of gravity of the parcel. More sophisticated systems could allow some topological attributes to be given to each parcel. The topology may be geographic or functional. For example, A is next to B, C and D, or A is connected by road to C and D etc.

The most important special case of the table is when the area parcels represented are uniformly shaped, preferably rectangular, thus forming a grid. In this case most of the spatial relationships between the data are maintained. The grid can be of any density to suit the problem. In the limit, where the grid is extremely fine, it is possible to identify the exact location of the boundary between different irregularly shaped area parcels and there is no spatial information loss at all. However, there is then much redundancy in storing the data, as parameters which vary slowly in space will have identical values stored in many grid elements.

The final case is the representation of the data in an analytical manner. Where a parameter is known to vary continuously, such as mean annual rainfall, a map may be prepared which shows the spatial variation of rainfall as a series of contours. In this case mathematical or functional relationships can be used to represent the variability of the parameter. This is highly efficient and can be of greater precision than a simple contour

map. The function, in addition to taking into account known spatial factors, can incorporate independent parameters which are known to physically affect the parameter being described, for example, altitude in the case of rainfall.

3.2 Data storage concepts

Having seen how data can be categorized into groups requiring different treatment for storage purposes, the general methods of data storage should be examined. Though methods vary between manual and computer methods, the concepts and terminology of storing data are largely identical. Thus, a conventional storage system will be described first, followed by a summary of the equivalent computer system to indicate those areas where the terminologies differ.

3.2.1 Conventional storage systems

In a conventional system data is stored in written form in pieces of paper or cards. The items of data relating to a single object, e.g. a place or an element (e.g. rainfall), will be put together on the same sheet of paper or at least on a collection of papers which are kept together. Each data group can be seen as a record of data about the individual object. If the record comprises a small amount of data, several records can be held on a single sheet of paper. Conversely a large record may occupy more than one sheet of paper.

Each data record will have some structure; each data item contained in the record is set out on the sheet of paper in a way that the item can be identified and its value given. The item name might be given followed by its value, or the item may be named only once and its value given by presenting it in a table of values. In the latter case the identification is made by position of the item in the table.

An item of information may be expressed as a string of text, such as a comment or remark; it may be a specific quality, its value denoted by a single descriptive word; or it may be the numeric value of a measurable parameter.

In order to assist with retrieval and updating it is convenient for collections of related records to be gathered into files. For convenience of reference, file records would usually be ordered in some way to permit rapid retrieval. A file may extend to several volumes if very large, with each volume occupying a physical storage device such as a folder, box, drawer, etc. A master file should give information about the location and contents of the main data files.

When data are required from such a system, the relevant file is searched for the piece of paper upon which the data are expected to be found. That sheet of paper is the smallest block of data that can physically be retrieved. Once obtained the paper can be scanned to obtain the necessary information. To locate this sheet of paper, all the sheets may be searched until the one required is found. Alternatively, the search may be facilitated by the existence of an index of file contents which points to the physical location of the data. This may be the exact location or a point from where to start looking.

Another aspect of data retrieval is the frequency with which a file or part of a file is used. If files are referred too often, for data retrieval or updating, it is desirable to have them, or the parts requiring updating, permanently to hand. Files not required so frequently can be stored at some remote location. These less accessible files are said to be archived.

To facilitate data storage and retrieval each record in a file must have at least one data item which serves as a unique identifier to enable that record to be distinguished from all others. By ordering the records in the file based upon this unique identifier, data retrieval can be made more efficient. However, if the records also need to be accessed on the basis of another data item which appears in the record, the efficiency is lost. One could save two or more copies of the file ordered by different data items in the records. It would be more efficient, however, to save one copy of the file ordered in some acceptable manner and also a series of indexes or lists which ranked the data according to the different data items. The indexes would be used to refer to records on the basis of the values of selected data items.

For example, in a soil inventory file, the major identifier for each record might be the soil survey site number. The data would be stored in ascending numerical value which would make data retrieval using this data item straightforward. If, however, it was required to know which soil samples were fluvisols, the first alternative would be to read every record, look at the position in the record where the soil type appeared, and select all records which had a fluvisol name or code. The second alternative is to have an index which listed for the data value fluvisol, each record containing a fluvisol soil. The index could have similar entries for each soil type. Other indexes could be made for any other data item which may be used as the basis for data retrieval, e.g. soil texture or structure. It will be seen that the computer offers in addition a third alternative which is to sort the entire file using soil type as the record identifier. This brings together all the records containing fluvisol soils, since their code would be identical, and these records could then be easily extracted.

Another simple example of indexes is the way in which a library provides an author index and a subject index. A book title is obviously its major identifier. However, the indexes enable the location of a given book to be found quickly by knowing either the author or the subject.

An efficient retrieval system enables records to be found on the basis of both their unique identifier and other selected data items within the record.

3.2.2 Computer storage systems

Computer storage systems are almost completely analogous to the manual system just described. The storage media is not paper or card but disk or tape. Related data items are collected into logical records, the space reserved for each data item being called a field. Data items which will be used as a basis for record retrieval are called keys. The most important key is the major record identifier. The units of storage media on which records are written are physical records rather than sheets of paper. Several logical records may occupy a physical record or vice versa. Similar records are

grouped together into files and several files are normally held in one physical storage device. In the case of the computer this is a tape, disk pack or diskette, also called a storage volume. Very large files may occupy more than one storage volume (multi-volume files as opposed to multi-file volumes). Each storage volume has a master file or catalogue containing the location and structure of the files held.

When data are retrieved by passing records one by one until the required record is found, this is called serial or sequential accessing. The alternative approach of creating indexes for each desired key and passing directly to any required record is called direct or random accessing. Files which are permanently accessible need to be held on mass storage devices, almost invariably disks, and are said to be on-line. Files not held in the machine are said to be off-line or archived. One feature of conventional storage systems used extensively in computer systems is the coding of data to reduce the bulk and improve the intelligibility of large volumes of data.

3.3 Coding of data

Two levels of coding need to be considered, the coding of information external to the computer, for data input or output, and the coding of data internally. When using a computer there is no reason why separate coding systems may not be used to gain efficiency. The advantage of coding is that the volume of data can be reduced and ambiguity is avoided.

A coding system is the representation of large amounts of information by a small amount. In effect the code is a pointer to a larger definition and becomes particularly efficient when the item being coded is observed very frequently. One detailed copy of the information is stored and is represented elsewhere by a coded value. These detailed copies can be grouped together by subject into "dictionaries".

For example, consider the field observation of soil texture "sandy clay loam". This may be transcribed onto a punching document or keyed directly into a data entry system encoded as "scl". This represents the external code. Before storage, however, the computer may convert this to the numeric code 5 by cross referencing with the dictionary entry for "scl". This internal code is more efficient in terms of data storage space than the external code. The dictionary entry might contain:

scl, 5, sandy clay loam, texture, ...definition.....

The definition is a text description of this texture for the purposes of the data base. For output, the external code or the original field observation may be used.

At a grosser level, instead of referring to an area having a whole series of physiographic, soil and land use characteristics requiring several pages of text to describe it, it may be called land unit type 15, where 15 refers to a detailed entry in a land classification dictionary.

In deciding upon coding systems for manual encoding and internal computer storage, two main points should be considered. In the case of manual coding it is better in general to use alphanumeric codes because operators can

remember combinations of letters and numbers and their meanings more easily than strings of numbers, especially if the letters are abbreviations or mnemonics of the original items.

In the case of machine coding it may be advantageous to use a code which gives the location in core memory or mass storage device where the full dictionary entry for the coded data item is stored. It may also be the location of a related record in another file. In these cases, numeric codes are to be preferred since they can represent storage addresses.

Whether or not a different set of codes is used for the internal and external representation of the data, the coding may be made completely transparent to the user, in that the single word WHEAT may be entered, or a collection of words, and these words will appear again in the output. It will not matter to the user how the data is represented inside the computer.

With conventional data-processing systems, in order to reduce input volumes and reduce ambiguities it has been common to let an operator encode the data for input to the system but let the machine decode them for output. This was because it was not possible, or it was extremely expensive, to put the machine intelligence and memory required for the coding task at the point where the data were being entered. The same difficulty was not experienced with output because the output device was always associated with the main computer. In general terms this situation has now changed and by use of microprocessors and remote terminals, machine intelligence can be placed at the point of data entry.

Despite the availability of these facilities there are still some good reasons to permit the human to encode the data for input. This is especially illustrated in the case of surveys, or other situations where it is difficult to foresee the complete range of code values that might be required. Also, there may be difficulty in spelling words and getting them out unambiguously. For example. a land use of BARLEY AND WHEAT may be classified for a particular application to be the same as WHEAT, BARLEY or just CEREALS. It is hard to write computer programs that will see these collections of words as meaning the same thing. It is often more efficient to let an operator decide on a particular code before data keying. The further benefit is, of course, that the key punch effort is reduced in both volume and intellectual effort. If coding is done at the key punch a better trained operator is required, and more sophisticated software must be developed and maintained.

An important feature of any coding system should be its relative permanence. Whether numbers, letters or mnemonics are used, mistakes are always made when a system is new. It is important to design coding systems which have the flexibility to accommodate new or expanded ranges of codes.

Each encoded data item requires a corresponding entry in a dictionary of allowable codes. In general each dictionary entry may give the external and internal codes, an abbreviated name, a full name, and a full description. If the same data item is known by other codes in other classification systems these may also be included for translation purposes.

An essential code for all kinds of data is to be able to know whether data are missing because they were not recorded intentionally or are missing because they were not recorded unintentionally.

4.0 Groundwater DATA

Groundwater data can be dealt with under three headings (as shown in Figure 3 : springs, wells and yields/costs. Well water level is still most often obtained by manual dipping techniques and usually at irregular intervals, producing an irregular time series format. If the wells are pumped, estimates of quantities abstracted may come from meters, from duration of pumping, or from quantity of power consumed. These estimates require knowledge of the pump specification and the pumping head, and may be performed by the processing system.

In addition to defining the surface drainage system of which the spring or well is a part, it is necessary to code the aquifer system(s). In the case of wells there is, in addition to the basic site description data, the full geological and hydrogeological logging of the hole.

It is probably better to save separate collections of data on wells, the details of discharge and water quality, the site location including licensing arrangements, and the details of the logging of the hole. Water level, discharge and water quality are time series data sets which can be stored in the standard formats presented elsewhere. Site location and logging are single sets of observations which do not change with time.

The natural resource data base system is well suited to storing geological and hydrological bore-hole data. Each horizon can be coded and the parameters of each horizon recorded. To devise the coding system for such a data bank is a large job and existing systems should be first reviewed to assess their suitability

However, a simple start can be made by noting against each bore-hole the location of reports on that hole in conventional form. This system would start by being an aid to data retrieval in a conventional library sense. For many kinds of data the coding and abstraction effort required to store the full, detailed information is simply not worthwhile for the number of occasions on which it would be used.

4.1 Station description data

Previous reference has been made to the two categories of data which are used to characterize a station: data which is fixed in time and data which varies in time, e.g. water levels and discharge. Generally the fixed data is input when the station is first incorporated into the data base, and the sets of observations made at the station are added periodically to the station time series file.

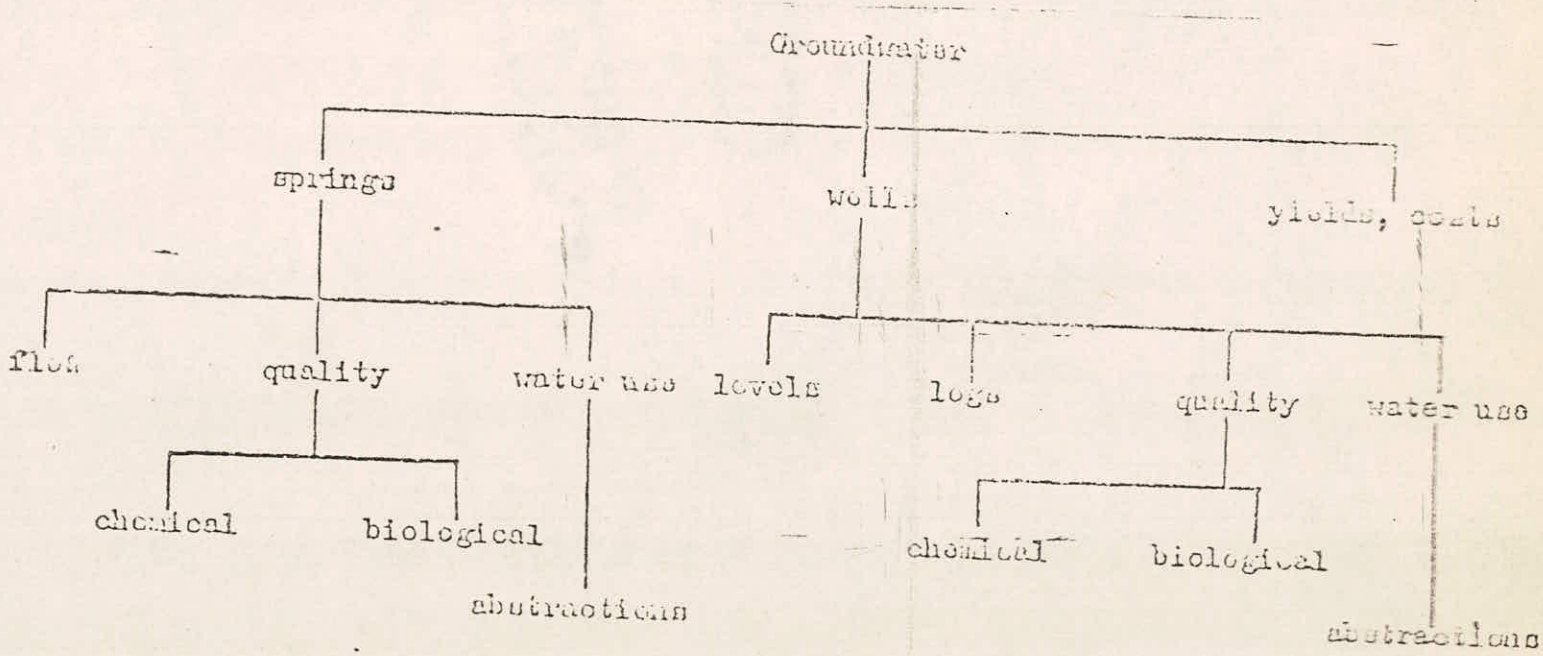

Figure 3 - The components of groundwater data

It has also been noted that this general view may be too simplistic for some types of hydrological station. For instance, whilst the prime objective of a flow gauging station is to record a time series of water levels, that is for many stations a second time series formed by the set of level-discharge relationships. In recognition of the time variation of some station characteristics, data structures of the type shown in Figure 2.6 have been utilized in several water data bases. The station history file is useful for all types of stations, whilst the level-discharge calibration file is required particularly for flow gauging stations. This structure allows the current station description and calibration files to be of fixed size and format, a useful feature when designing retrieval programs. The station description file contains data on the current instrumentation and time series data location and format and if this needs revision it is first copied to the station history file then replaced in the station description file by the new data.

For the purposes of discussion all the data except the specific station time series, e.g: daily rainfall values, water levels, and water quality values will be classified as station description data.

Whether conceived as a single file or broken down the variables needed to describe the location, purpose, equipment, administration, and operation of hydrological stations can be readily identified. Table 2 summarizes these data, but also gives examples of station data specific to some of the hydrological variables identified in the previous sections.

Site description data files are essentially as found in any manual system, although more emphasis is placed on the location and format of other relevant data. For example, computerized hydrological station files need to contain explicit references to the disposition elsewhere in the computer of the associated time series files, or the dictionary files which convert codes for station, instrument, and analysis types, watershed names, data reliability etc.

REFERENCE:

"Guidelines for computerised data processing in Operational Hydrology and Water Management", Joint FAO/WMO publication, WMO-No.634, 1985.

TABLE 2

Typical hydrological station description data

1.	Type of station	– river level, bore-hole etc.
2.	Station number	
3.	Grid coordinates	– latitude, longitude, or Universal Transverse Mercator projection (UTM)
4.	Description of site location	– how to get there, details of where to find surveys and full original details
5.	Hydrological location	– hydrological reference. Drainage area and/or aquifer(s)
6.	Details of each instrument at station	– type – manufacturer – serial number – pointer to rating/calibration curves – limits of calibration – time and measurement scales of data recorder – date installed – date last serviced – frequency of servicing
7.	Details of licensing	– name of licensee – address – consent conditions

TABLE 2 (contd.)

	- amount to discharge (-) or abstract (-) conditions may depend on time of year or flow condition (level or discharge)
	- water quality - range of values allowed for each coded parameter, for each discharge condition
8. Structure of the time series input data formats	- for each kind of input record, the position, format and units of each data field
9. Structure of the time series storage data formats	- format (Figures 2.1 and 2.6) - number of data values per record - are multiple parameter data values identified by position or by parameter code plus data value?
10. Physical organization of time series files	- which disks disks/tapes etc. - for what period is data available (including missing segments)
11. Details of environment	- data include exposure of site, altitude, etc.
12. Details of cross-sections	- for river sections only, includes datum of location of verticals and level gauge. This may contain only the latest section or it may contain a selection of previous sections. If frequency of survey is high it may be necessary to create a separate time series file for this category.
13. Bore-holes	- a basic minimum set of well data includes the depth, top of well datum, diameter schedules and which aquifer(s) is tapped (coded). - reference to detailed manual or computer records to facilitate rapid access.

TABLE 2 (contd.)

- coded geological succession and analytical parameters
- pump test data in standardized form
- constructional details

NOTE: The drainage area or aquifer codes must enable the discharge to be logically related to the basin or aquifer. In the case of individual aquifers x-y coordinates and definition of the pumped aquifers are needed. In the case of surface discharge the site should at least be topologically ordered.