A DATA BASE DESIGN FOR A SECONDARY SCHOOL GUIDANCE INFORMATION SUPPORT SYSTEM

By

EDWIN FORSBERG

Bachelor of Science
Milwaukee School of Engineering
Milwaukee, Wisconsin
1957

Master of Business Administration
Oklahoma State University
Stillwater, Oklahoma
1966

Submitted to the Faculty of the Graduate College
of the Oklahoma State University
in partial fulfillment of the requirements
for the degree of
DOCTOR OF EDUCATION
May, 1970

A DATA BASE DESIGN FOR A SECONDARY

SCHOOL GUIDANCE INFORMATION

SUPPORT SYSTEM

Thesis Approved:

Kenneth E. Wiggins
Thesis Adviser

Fred Black

Donald R. Gamble

John T. Hampton

D. Durham
Dean of the Graduate College

ACKNOWLEDGMENTS

I would like to express my deep appreciation to all who assisted me in the preparation of this dissertation. Special thanks goes to the chairman of my advisory committee, Dr. Kenneth E. Wiggins, whose constant encouragement and assistance led me to the completion of the entire degree program. I wish also to thank the members of the advisory committee, Dr. John D. Hampton, Dr. Rondal Gamble, and Mr. Fred Black, for their advice and leadership.

Thanks also goes to Dr. Tommy L. Roberts, Southern Methodist University, who provided very important advice pertaining to the substantive input to the study.

Thanks to Mr. Earl Hamon and Mr. Wayne Richardson of the Bartlesville public schools for their cooperation in furnishing personal time, space facilities, and descriptive input to the data base design.

I express a deep gratitude to my wife, Joie, and my two daughters, Marianna and Aimee Leigh, whose sacrifices made the total degree program possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER I

THE PROBLEM

Introduction

With the twentieth century has come numerous problems associated with the access and utilization of basic information relevant to individual achievement in many areas of life. Along with these problems has come the need to improve upon conventional systems related to the acquisition of information necessary for individual guidance and self actualization. The need for new and better ways of helping man improve his lot in today's complex society cannot be underemphasized. The purpose of this study is to design a data base which can be used as an exemplar for a computerized guidance information support system.

Statement of the Problem

Therefore, it seems both necessary and timely that increased attention should be given to designing systems that utilize present day capabilities in meeting these needs. An adequate model has not been developed for this purpose and this study attempts to close this gap.

Assumptions

There were certain basic assumptions made in the design of such a data base. Among these are the following.

1. There is a body of information that is vital to facilitating the guidance function in the secondary schools.

2. Information is hierarchial in terms of its importance in the educational environment specifically as it relates to the guidance function.

3. A digital computer can be used to retrieve and process information faster and more accurately than any present day human system.

Questions

Several questions were answered while attempting to design the data base model. Among these are:

1. What constitutes an adequate data base of student fact data for a secondary school system?

2. How should the data records be structured to facilitate retrieval and processing?

3. How should the data records be ordered within the data file to facilitate retrieval and processing?

4. How many data files should be built into the data base?

5. How should the data files be organized within the data base?

6. What is the role of the storage medium in the organization of the data base?

7. What will be the file maintenance procedures for the resulting data base design?

Procedure

The procedures used in designing the proposed data base model were (1) a review of the literature related to data base design and information systems, (2) an analysis of the information needs in the Bartlesville secondary schools, (3) the determination of how information is used in facilitating the counseling function, and (4) the development of a student data base model and an outline of file maintenance procedures for a synergistic information support system.

The related literature was found in journal and magazine articles, books, and IBM reference manuals. These sources were gleaned for facts pertinent to the usage and design of real-time information systems. It was discovered that the utilization of real-time information systems in guidance and counseling are rare and discussions of the data base design are practically non-existent.

The information needs of the guidance function were determined by several visits to the school system and meetings with the counselors and other administrative personnel in the school system. The primary purpose of the visits was to determine the present system's elements and their relationship to each other which consequently depicted how the information was used. The purpose of the meetings was to determine the substantive input to the data base. The results of these visits and meetings are discussed in Chapter IV.

Delimitations

The scope of this study was limited to the following.

1. A data base consisting of a student fact data designed for TGISS (Total Guidance Information Support System) for the Bartlesville secondary schools.

2. The utilization of an IBM 360/50 computer linked to remote 2260 graphic display terminals.

3. General user requirements as identified by a team of guidance counselors and judges.

4. User requirements for guidance purposes.

5. A data base designed to facilitate the guidance function.

Teleprocessing Environment

TGISS is a real-time information storage and retrieval system specifically designed to meet the needs of the guidance function of the Bartlesville secondary schools. The system is a combination of man and machine. Emphasis is placed, however, on man who uses the machine only as a tool to expedite the counseling task.

The "tool" exists to provide timely information about the student. This information is in a location remote from its point of display. Specifically, the information is located in the disk files of an IBM 360/50 computer located on the campus of Oklahoma State University, Stillwater, Oklahoma. Once the counselor enters the necessary "command," the system responds by retrieving the data from its files, formulating the data for display, and sends the formatted data over telephone lines to the remote display terminal. Thus, the machine element of the environment consists of computer and storage hardware, a communications network, a remote display station, and system software.

Remote Terminal

The remote station consists of an IBM 2260 video display much like the ordinary TV screen. The major difference is that it cannot display pictures--only characters. The screen capacity of the 2260 screen to be used in TGISS is 960 characters arranged into twelve lines of eighty characters each. A typewriter-like keyboard is an integral part of the 2260 display station which provides the counselor with a facility for entering data into the system as corrections or changes.

The 2260 display station is cable connected to a control unit, the IBM 2848, which provides the following functions.

1. Translate and store the data coming from the telephone lines.

2. Generate the display characters.

3. Timing and control logic to synchronize the internal and display operations.

Up to eight 2260's can be attached to a 2848. TGISS ultimately calls for six 2260's, three in each of two high schools. Since the maximum cable length is 2000 feet, there will be two 2848 control units.

An IBM 1053 printer will be attached to each control unit. The printer provides a hardcopy from either the computer or the 2260 keyboard. It will be used to provide the user with a printed copy of the 2260 video display.

Computer and Storage

The central computer facility at OSU consists of an IBM 360/50 processor, an IBM 2314 disk storage device, and an IBM 2701 transmission control unit. The 360/50 is a high speed processor with 256K bytes of high speed storage and one million bytes of low speed storage. The 2314 consists of eight disk drives with a capacity of 29.17 million bytes each giving a total of 233.4 million bytes of mass storage. TGISS will require disk storage space equivalent to about one drive. The central computer is supported with IBM's OS operation system operating in the MFT multiprocessing mode.

Communications Network

The communications link consists of a full-period leased line operating in half-duplex (send and receive, but only one way at a time). At each end of the line is a modem which is needed to convert digital data at the sending end prior to transmission over the telephone line and to convert back to digital data at the receiving end. This equipment is supplied by the Southwestern Bell Telephone Company.

System Software

System software refers to the computer programs which perform the tasks required by the fact retrieval of TGISS. The software consists of essentially the three parts (1) file creation, (2) file maintenance, and (3) file search and processing. The first two parts are written in PL/1 and the third in FASTER.

The function of the file creation programs is to validate certain input data and to initially build the data base. The file maintenance programs are designed to add, delete and alter the contents of the data base. Finally, the search and process programs serve to access and format selected student data elements into aggregates for meaningful display.

Clarification of Terms

Activity, record: A term which indicates that a master file record is used, referred to, or altered.

Byte: A group of eight binary digits usually operated on as a unit. Each of the alphabetic characters is coded with an eight bit byte.

Data Bank: A group of data bases which may or may not be related.

Data Base: A group of data sets or files which are usually related to each other.

Data File: A group of related records.

Data Group: A group of related data elements.

Data Item: A group of related data elements or data groups.

Data Set: Same as data file.

Element: A group of bytes, or characters, which is treated as a whole. This grouping is also called a field.

FASTER: Filing and Source Data Entry Technique for Easier Retrieval, a language of inquiry macros designed to facilitate information retrieval.

File Maintenance: The process of keeping a data set up to date by adding, deleting, or changing data.

Half-duplex: The transmission of data over a single communication channel in both directions, but one way at a time.

Hard copy: A printed copy of machine output in a visually readable form. Examples are printed reports, listings, documents, and summaries.

IBM: International Business Machines, Incorporated.

Modem: A contraction of the two words modulator and demodulator. A modem is a communications line device which adapts the telephone and data channels.

PL/I: Programming Language/I, a high level programming language.

Query language: A programming language designed for the retrieval of information facts from an existing data base.

Real-Time: Pertaining to the actual time during which a physical process takes place.

Record: A group of related data items.

Software: A set of programs, procedures, rules, and related documentation concerned with the operation of a data processing system.

Storage Medium: The material on which data are recorded, for example, paper tape, magnetic tape, cards.

Student fact data: Factual data about the student such as name, age, grades, health condition, etc. which are gathered and stored for future retrieval during the counseling session.

Subroutine: A routine that can be part of another routine.

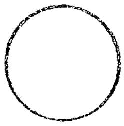
Teleprocessing: An IBM registered term referring to systems which transmit data from one remote point to another in the course of processing.

TGISS: Total Guidance Information Support System as funded by USOE Grant No. 7-8-005685-(056).

Third generation computer: Computing equipment which features microminiaturized circuitry, thin-film memory, and other advanced electronic technology which make possible operating speeds measured in billionths of a second. The third generation is contrasted with the first generation which utilized vacuum-tube circuitry, and the second generation which utilized solid-state (transistors, for example) devices.

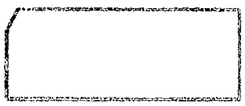
Volatility, record: A term which refers to the addition and deletion of records from a file. A volatile file is one which has a high percentage of additions and deletions.

Definition of Flow Chart Symbols

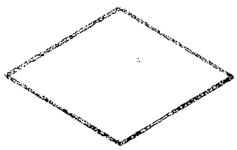


Start or Stop

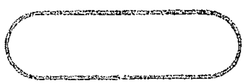
Processing Function



Card Input/Output

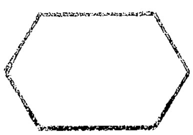


Decision Function

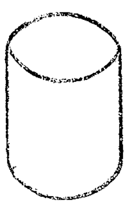


Connector

Connector



Subroutine



Disk Input/Output

CHAPTER II

BACKGROUND OF EDUCATIONAL GUIDANCE SYSTEMS

History of Data Processing

Contemporary educational guidance systems are a direct result of the developments of the electronic data processing equipment. The equipment came into existence as a result of a need to process vast amounts of data produced by the United States census. The processed data provided meaningful information to its users. Guidance counselors also have large amounts of data to contend with which can be processed on the machines conceived by man some one hundred and fifty years ago. Hence, the first part of this chapter will discuss the "hardware" developments leading up to the educational guidance systems.

Whitlock (47) defines data processing as a series of planned actions and operations on information designed to produce a desired result. A data-processing system consists of procedures (software) and devices (hardware). Data processing systems can (1) be entirely manual, (3) use key-driven machines, or (3) use the so-called automatic data processing devices. The greatest potential for education lies within the third type of system. There are two types of machines found in the automatic system. First, the punched-card, or unit record, machines which are electro-mechanical in design and second, the electronic data processing machines or the electronic stored program computer (47:10).

The punched card concept dates back to 1801 when Joseph Marie Jacquard, a Frenchman, invented the punched card textile loom. This machine utilized punched cards to guide a machine which performed intricate weaving (1).

The next significant development came about one hundred years later when Dr. Herman Hollerith developed punch-card equipment for the purpose of tabulating census data. His first machines were introduced in 1887, and they were used on the 1890 census. These machines cut the processing time down from the seven and one-half years required of the 1880 census to two and one-half years for the 1890 census not considering the twenty-six percent increase in population (6).

Dr. Hollerith resigned from the Census Bureau to form his own company in 1903. In 1905 the Census Bureau hired James Powers to develop additional equipment for the 1910 census. From this point on, Powers and Hollerith paralleled the development of machines such as the key-punch, sorter, collator, reproducing punch, and tabulator (6). The concept of these machines was relatively simple. Data were punched into a card using a specific machine readable code. The data were recorded using a standard format and this "position" of the data had meaning to the data processing system. Each machine in the system had a specific purpose. For example, the sorter could arrange the cards into desired groupings, and the tabulator could summarize by counting and totaling the data. Because the card contained a "unit" of data, the card became known as the <u>unit record</u> from which <u>unit record data processing</u> developed. Our present day unit record equipment came from the early work of these two men. The contemporary unit record hardware produced by IBM (International Business Machines) evolved from the innovations of Hollerith and the hardware of Remington Rand evolved from that of Powers (1).

Although the unit record principles expanded man's ability to process data, it lacked flexibility and was limited in speed and calculation complexity. Only the simplest of operations could be performed, that is, addition and subtraction. An attempt to increase the computation ability of the unit record equipment was made by adapting the calculator to a mechanical automatic controller and the unit record input/output devices. The purpose of the controller was to perform the steps normally done by the human calculator operator. The data were fed into the machine from punched cards; the calculation steps were performed by the controller which was programmed from a wired board; and, the results were punched into cards.

The mechanical calculator and controller, and the wired board program, proved to be too slow and too inflexible. The mechanical controller/calculator limited the speed and the wired program made it difficult to change the desired sequence of operations. These limitations were reduced by evolving from the strictly mechanical controller/calculator to high speed all electronic circuitry, and from the wired program to the internally stored program. The first all electronic calculator was part of the ENIAC (Electronic Numerical Integrator and Calculator) developed between 1942 and 1946 at the University of Pennsylvania by Drs. J. Presper Eckert and John W. Mauchly. This machine still lacked the flexibility of programming because of its externally wired board, but its computation speed was increased considerably; the ENIAC could perform in thirty seconds what a human could compute in about twenty hours.

In 1945, John Von Neumann proposed a solution to the wired board problem by suggesting that the program (sequence of steps telling the machine what to do) be stored in the computer's own memory. As a result of Von Neumann's proposal, the ENIAC was converted to a partially stored program computer in 1947 and his own system, the EDVAC (Electronic Discrete Variable Automatic Computer), was completed in 1950. The first fully stored program computer was the EDSAC (Electronic Delayed Storage Automatic Computer) developed in England in 1949.

The first commercially applied general-purpose digital computer was the UNIVAC I developed by the designers of the ENIAC at the Sperry Rand Corporation. The first UNIVAC delivery was to the Bureau of Census in 1951. IBM marketed its IBM 650 about the same time (1).

Historical Developments in Guidance

Data processing in public school education dates back to the late 1920's and early 1930's. At this time the larger school systems of cities such as Detroit, Baltimore, Cincinnati, Philadelphia, and St. Louis began using punched card systems for routine administrative tasks as payroll and fixed assets accounting. The smaller schools had to rely on the service bureaus, which were being established to serve private industry, since they could not justify an installation of their own. The services provided by the service bureau were similar to those implemented by the larger schools, but the service bureau took the initiative in developing systems for attendance, class scheduling, test scoring and recording, and personnel records (18). Other uses of unit record data processing include: annual census records; registration; grade reporting; honor roll and failure lists; grade distribution;

curriculum planning; supply requisitioning; school expenditure accounting; voucher check writing; check reconciliation; receipt accounting; and encumbrance accounting (47). There are many more uses as indicated by Anderson (3) who lists 100 applications for school data processing.

As for guidance, per se, the initial impact of unit record and electronic data processing came in the late 1950's (33). It wasn't until 1950 that counseling and guidance came into its own. During the period 1950-1957, counseling and guidance began an ". . . attempt to develop a systematic theory based on a philosophy of education; and from 1957 to the present--developing various schools of thought" (40:7). Included in this systematic approach to guidance were the routine tasks of attendance, class scheduling, test scoring and reporting, and student personnel records. Thus, it was natural that the administrators call on the counselor to develop data processing systems for guidance. The "trial and error" development of systems needed to expedite some of the guidance functions continued until 1967 when "The use of computers in the mechanical and record-keeping guidance functions noted is no longer experimental" (33:23). The computer began to "catch on" in guidance around 1964.

One advocate of the "computer systems" was Cooley (14) who, as a result of a plea from psychologists that the teacher and counselor should use and interpret test scores in relation to all data available on a given student, proposed a computer-measurement system which would monitor the student variables longitudinally and perform a multivariate analysis as a basis of sound interpretation of test data. "By a systematic examination of several different types of data, a variety of potential problem cases can easily be uncovered today among a very large group of students. A computer-measurement system could allow the early identification of potential problems, soon enough to do something about it" (14:164). Once the problem student is discovered, he would be taken through a set of "programmed experiences." The sequence of experiences in the program would provide such things as (1) an exploration of the world of work yielding exposure to the jobs open to this kind of student, the job training requirements, etc., (2) assistance with college placement including the matching of the characteristics of the student and college, and assistance with selection of an area of study, and (3) talks with a counselor. Adjunct to the programmed experiences would be (1) an analysis of student achievement, (2) diagnostic functions, (3) systems analysis of curriculum practices. Cooley's proposal seems to be the foundation for the contemporary studies to be discussed in the Related Research section of this chapter.

One computer system developed as a result of Cooley's suggestions is that of the Covina-Valley Unified School District, Covina, California. Smith (44) describes the system as one which monitors the course selections made in light of the student's postgraduate plans. The system utilizes the student's GPA to measure the success of his plans along six dimensions of the so-called "alert conditions."

A career planning system for college students at the State University of New York is in its early phases of development with intentions for eventaul computerization. The system presently utilizes a very low level of unit record data processing to assist the upperclassmen with making decisions about his post-graduate plans (2). This system can easily be converted to a computer system when such equipment becomes available to the university. Although this is not a system which exists in the public schools, it is mentioned here because it has all the characteristic needs of career planning at the high school level and it can easily be emulated for use in the public schools.

Another computerized guidance system which could have substantive meaning to the high school guidance program is the Computer Managed Instruction (CMI) system (12). This system is being developed at the New York Institute of Technology under the direction of Dr. Bertram Spector. The purposes of the system are:

1. The development of course and curriculum objectives.

2. The development of student records and profiles which reveal not only the student's achievements but also his difficulties and patterns of difficulties.

3. To suggest the best strategies for coping with the student's individual characteristics and for plotting his academic development.

4. The development of actual materials of the course and curriculum.

5. To assess whether or not goals have been reached.

6. To provide "feedback," or information which will enable the system to improve itself.

7. To organize personnel and facilities required to support these activities.

Systems which make pertinent information available to the counselor but are not designed specifically for the guidance program are the statewide information centers (8,35,42). The primary purposes of these systems are to avoid costly duplication of design and implementation, and to standardize the syntax and semantics of educational information.

The systems mentioned above have been designed and put into use as a solution to immediate local problems and do not represent a "total approach" to the problems of a guidance program. However, "Probably computers' most important impact on guidance and counseling is in research and evaluation" (33:23). One aspect of the use of the computer in guidance research is to use simulation techniques ". . . to facilitate the clarification and testing of counseling theory. . ." (21:47), that is, pure research. Another aspect of the research function is to facilitate the clarification and testing of the application of counseling theory, that is, applied research. The project, TGISS, for which the data base is to be designed falls into the category of applied research. The next section discusses the research projects which are related to TGISS and it should be noted that they have aspirations of solving the very broad problems of guidance through the use of systems commonly called "computer-assisted counseling."

Related Research

Based on the traditional involvement of the counselee and counselor in the decision-making process, Bohn and Super (9) specify several roles of the computer as a tool for making this process more effective.

1. As an information source for the counselor, giving him data about; his client, educational resources, the academic promise, the world of work, career patterns and prospects, etc. The counselor and client together retrieve data for input to the decision-making process.

2. As an information source for the student, giving him information which he can ponder and use in his own way. The student alone retrieves data for input to the decision-making process.

3. As a "client and computer interaction" system. The student

and machine converse with each other. "The student is assisted as he moves forward toward a decision of his own making; he is given all the information the machine can provide and all the assistance counselors and programmers can put into an interaction system for the purpose of making a decision as knowledgeable and helpful as possible" (6:30).

Loughary and Tondow (34) classify computer-assisted counseling systems on three levels:

1. Systems which perform the traditional <u>information processing</u> tasks and are essentially data processing tools for the counselor. These tools aid in the performance of such processing tasks as grade recording, attendance recording, test scoring, etc.

2. Systems which are <u>substitute for counselors</u> that perform, in addition to the information processing of level 1, some of the counselor's tasks such as the scheduling of courses.

3. Systems which are <u>substitute counselors</u> which take on characteristics found in human counselors such as privileged communication or privacy.

Regardless of the role or level, the basic educational need is one of more appropriate information and fewer unsubstantiated opinions as input to the decision-making process. This fact is clearly pointed out in the literature related to computer-assisted counseling.

One of the earliest exploratory designs of a man-machine counseling systems for secondary schools was developed by John Cogswell and Donald Estavan of the Systems Development Corporation, Santa Monica, California (11). The project attempts to simulate the counselor's behavior in the areas of the cumulative folder appraisal and the student planning interview. A simulation model was derived from careful analysis of recording made by one counselor of twenty counseling interviews with ninth-grade students. "The automated cumulative folder appraisal system accepts as inputs the data in the cumulative folder--grades, aptitude test scores, parent's occupations, and so on. The program analyzes these data, applying the programmed 'rules' abstracted from the counselor's verbal behavior, and selects output statements such as the following:

> ... 'Student is a potential dropout.'
>
> ... 'Low counseling priority. No problems apparent.' ..."
>
> (11:161-2).

The interview takes place with the student seated at a teletypewriter terminal. The result of the interaction with the "simulated counselor" is a tentative plan of study for grades 10, 11, and 12.

An integral part of the automated interview is the student-information system consisting of (1) a data base stored on magnetic tape, (2) an information retrieval system for student print-outs, and (3) a set of programs (models) which appraise the data and which conducts the study planning.

A study closely related to the Cogswell-Estavan project is one conducted by Tondow and Betts (46) of the Palo Alto Unified School District, Palo Alto, California. This counseling project utilizes a form of computer-based instruction to help the ninth-grade student plan his courses for the ensuing semesters. "The student taking the dialogue learns something about himself and about his chances for success under various conditions that are open to him as options" (46:216). This dialogue is a student-computer conversation where the student begins by typing his identification number after which the computer comes back with the student's name and asks for a verification. If the student name and identification are legitimate, the computer presents the student with a list of courses that he had taken as indicated by his last report card. He is then asked if he is having any problems with any of the courses listed. If so, he is asked to indicate the nature of his problem by typing it on the typewriter terminal (a list of diagnostics is eventually sent to the counselor). The student-computer session continues with a non-prescriptive dialogue which presents the student with probabilities associated with his GPA and choices related to course selection. The purpose of the project is to provide the student with the information needed to make good decisions.

The remaining computer-assisted counseling systems emphasize vocational decision making. One such system is the Education and Career Exploration System (ECES) developed by IBM, Advanced Systems Development Division, Frank J. Minor, in cooperation with counseling psychologists at Teachers College, Columbia University, Donald Super and Roger Meyers (9,38).

ECES is designed to serve "...individuals in the late growth and exploration stages, in junior and senior high school, and through the first years of post-high-school training or work experience when people seek to implement their self concepts" (9:30). Based on the task of relating the understanding of the individual's environment, ECES incorporates three phases which are crucial to the growth and exploration stages; (1) orientation to the world of work, (2) educational orientation, and (3) training and job search.

ECES uses a computer to store all kinds of educational and occupational data such as (1) working conditions of jobs, (2) type and amount of education needed for particular job types, (3) grades and standard test scores needed to enter specific colleges, and (4) data about the student who interacts with the system.

The student uses computer-based instruction methods in a conversational mode to learn about the world of work and then he makes inquiries about vocations. The information system provides a list of jobs which satisfy his preferences. The compatibility between the student's characteristics and the job characteristics is checked and suggestions made. Another list of jobs may be suggested by the computer.

Another large effort oriented toward the solution of some of the problems related to decision making in the world of work is a project called Information System for Vocational Decision (ISVD) being developed by Tiedeman, et al., at Harvard (27). "The major objective of ISVD is to improve vocational decision-making through the use of a computer-based guidance system" (27:2). Tiedeman considers data to be facts and information to be interpreted facts; and it is the task of the "guidance machine" to enable the student to transform data into information. "This is done by teaching him to interpret the data in the light of his own knowledge, experience, and intention, so that his organization and use of the data represents his own personal relationship to them in the process of decision-making" (27:2).

In ISVD, the counselee and the computer system interact in three processes of career development: (1) The exploration process which begins when the counselee becomes aware that a decision must be made, (2) the clarification process where the counselee improves his self-understanding and attempts to remove unrelated alternatives pertaining to the decision, and (3) the review process where the decision maker systematically evaluates his decision in light of the information available (9:31).

Impellitteri's (23,24,25,26) Computer-Assisted Career Exploration (CACE) system was developed with three purposes in mind.

1. To provide an occupational information retrieval system.

2. To create a process where an individual can discover his own framework of occupational structure.

3. To provide a simulated environment in which the individual can gain experience, through simulated practice, in operational strategies related to occupational choice.

The mechanical elements with which the counselee interacts are a typewriter terminal, tape recorder, and slide projector, all under computer control.

"Information related to certain students' abilities, preferences and educational plans are stored in the computer before the student begins the interactive phase" (26:vii).

An orientation to the system provides the student with a list of 40 coded occupations. During the student-computer dialogue, he is asked to select and enter one of the 40. Upon doing so, the machine system responds with a brief description of the occupation. Next, he is asked if he would like additional information about the job. If so, four operations are set into motion.

1. Discrepancies between the job requirements and the student's characteristics are typed out.

2. A two-minute interview with a worker is played.

3. A slide is projected depicting a typical task of the job.

4. A 150 to 200 word job description is typed out.

At any point in the interview, the student can ask the computer to select a job list which is based on the student's aptitude-preference profile. It lists those occupations compatible with the student's characteristics. The occupation file is limited to forty.

Flanagan's (17) project PLAN is similar in that it includes a computer system which monitors the student's development and improvement, and a planning program with three aspects.

1. Provides the student with occupational data.

2. Gives the student information on where he stands in relation to others.

3. Gives the student practice in sound decision making in a gaming situation.

Another career guidance project initiated by Johnson (29) and apparently being currently developed by Youst (31) is the Rochester Career Guidance Project, City School District, Rochester, New York. This project calls on three agencies to fit together a system to increase the effectiveness of school counseling: A public school system; a state employment service; and, a private corporation (Eastman Kodak). The system was proposed to have at least the following elements:

1. Student descriptive data to be stored in computer files,

2. Occupational and educational data to be stored in computer files,

3. Audio-visual support information,

4. Audio-visual display devices,

5. Counselor-student users,

6. Computer programs, and

7. Dialogue capability between the machine system and the users.

Finally, the project of JoAnn Harris (13,20) called Computerized Vocational Information System (CVIS) is being implemented at the Willowbrook High School, Villa Park, Illinois. CVIS makes accessible to the student specifics on training and academic preparation of over 400 vocations. As in the previous studies, the system contains basic characteristic data about the individuals who intend to use the system. The student (a sophomore, junior, or senior) makes contact with the vocational files through a typewriter keyboard attached to a TV-like terminal. A dialogue built into the system describes eight occupation areas in general terms and then asks the student to select an area of interest. The machine then produces a list of all jobs in the area on a typewriter output attached to the terminal. From this list the student makes a selection and asks the machine for a fifty-word description which is presented on the TV screen. He can also request an additional 300-word description if desired. The objectives of the system are (1) to facilitate vocational decision making and (2) to make occupational and student data readily available to the counselor.

The review of the literature clearly points out that "Experts generally agree upon certain basic elements needed in any effective guidance program. Two sets of variables stand out as essential to each activity aimed toward providing good guidance services for the student" (19:171):

1. Data concerning the individual--abilities, interests, aptitudes, health, special experiences, and sociological background.

2. Information about the areas wherein he must make choices and plans. The need for sound guidance, moreover, is heightened by the current national shortage of manpower. There is a direct and

significant relationship between the effectiveness of guidance services and the directions high school graduates take or will be taking in pursuit of occupational goals.

CHAPTER III

RUDIMENTS OF INFORMATION ORGANIZATION

Introduction

One of the basic assumptions of this study is that a digital computer can be used to retrieve and process information faster and more accurate than any present day human system. This assumption is based on the fact that "The digital computer has, for some time, been playing a key role in the field of information science and technology," and that "Large files of data stored in digital form have been the primary factor in all applications of this technology" (37:123). The purpose of this chapter is to present some of the concepts of the organization of information as it relates to the design of a data base which will be a component of the guidance counselor's data management system.

Information Structure

The organization of information for retrieval purposes implies a highly structured relationship between its elements. For example, the "Organization of documents for retrieval is based on the assumption that it is possible to describe the content of documents adequately by assigning subject tags to them" (4:18). This assignment of "subject tags" is called <u>coding</u>, or the assignment of symbols to information. Coding occurs at several levels and moving to a higher level involves the combination of codes of the lower levels. Levels of coding in the

English language, for example, are character, word, phrase, sentence, paragraph, etc. The hierarchy of information coding to be used in this study is from lowest to highest, the bit, byte (character), element, data group, data item, record, data set (file), data base, and data bank.

The word bit is a contraction of the two words binary and digit. The binary digit, or bit, is the smallest unit of information of the binary machine (digital computer) language. The bit symbols are 0 and 1.

The byte is a group of bits usually operated upon as a unit. The number of bits in the group is arbitrary, e.g., a 6-bit or 8-bit byte, but once selected the number is constant. The bit combinations of the byte usually represents symbols from the natural language. For example, the 8-bit EBCDIC (Extended Binary-Coded-Decimal Interchange Code) bit combinations for the English alphabet are

EBCDIC Byte	LETTER
11000001	A
11000010	B
11000011	C

An element is a group of bytes, or characters, which is treated as a whole. The data element is comparable to the word in the natural language. The term element is synonomous with the term field.

A data group is a collection of elements which are related in some way.

A data item is a set of one or more elements or data groups containing related information. For example, several repeating elements would be required to record the absences for a thirty-six week period.

The thirty-six elements taken as a group is a data item.

A record is a group of data items which are related to each other. For example, a student record could include his identification, name, address, age, sex, or any other field considered appropriate for a given situation. The organization of fields into records is a basic design problem of retrieval systems. The set of rules used to combine fields into records is called syntax (36).

A data set, or file, is a collection of records grouped for a specific purpose. Sippl (43:88) defines a data set as a "Unique combination or aggregation of data elements." The records in a data set are usually related as to information content, e.g., a health record data set may contain records of diseases and their dates of occurrence. The terms data set and file will be used synonomously throughout this study.

A data base is a collection of one or more data sets or files. For example, a student data base may consist of a set of files covering personal data, course data, teacher data, health data, transcript data, etc.; or the data base may be a single file containing all of these data. If the data base consists of more than one data set, a member file either uses data fields belonging to another member or contains a direct reference to another member file (36).

A data bank is a collection of data bases which do not necessarily have a cross-reference relationship with each other. This does not preclude, however, that they cannot be related. The basic notion is that some underlying body such as government, education, banking, etc., collect and process diverse files for the purpose of providing a general service, overall control, a source to determine overall trends, or any other transcendent service. The repository concept of the data bank offers advantages of (1) standardized data collection, storage and dissemination and (2) the elimination of redundant variables and data sets (45). An example of a data bank is that of EDUNET of EDUCOM (22). EDUCOM is an acronym for educational communication and EDUNET is its corresponding network. EDUNET is presently a conceptual nationwide network linking together data that "...will permit the instant sharing of audiovisual and other sources, of books, documents, research results, and even the time of professors" (22:352).

Information Retrieval

The literature delineates two kinds of retrieval systems, viz., document retrieval and data retrieval (32,36,41). Document retrieval involves the display of a long symbol string, such as a book, and/or a reference to its location. Data retrieval involves a display of short symbol strings, such as words or numbers (a student's name and address, say) which are specific answers to questions. This study is concerned with data retrieval.

In order to put the elements of a data base design into its proper perspective, it seems appropriate to discuss briefly the information retrieval process. Figure 1 depicts this process as it applies to this study. The process begins by collecting the data on the desired variables about the student, teacher, courses, etc. The next step would be to code the variables at various levels such as the byte (character), field, and record. Then comes the problem of classifying the variables into data sets. This step answers the question "How many data sets?" Each record in the data set has associated with it an index, or key,

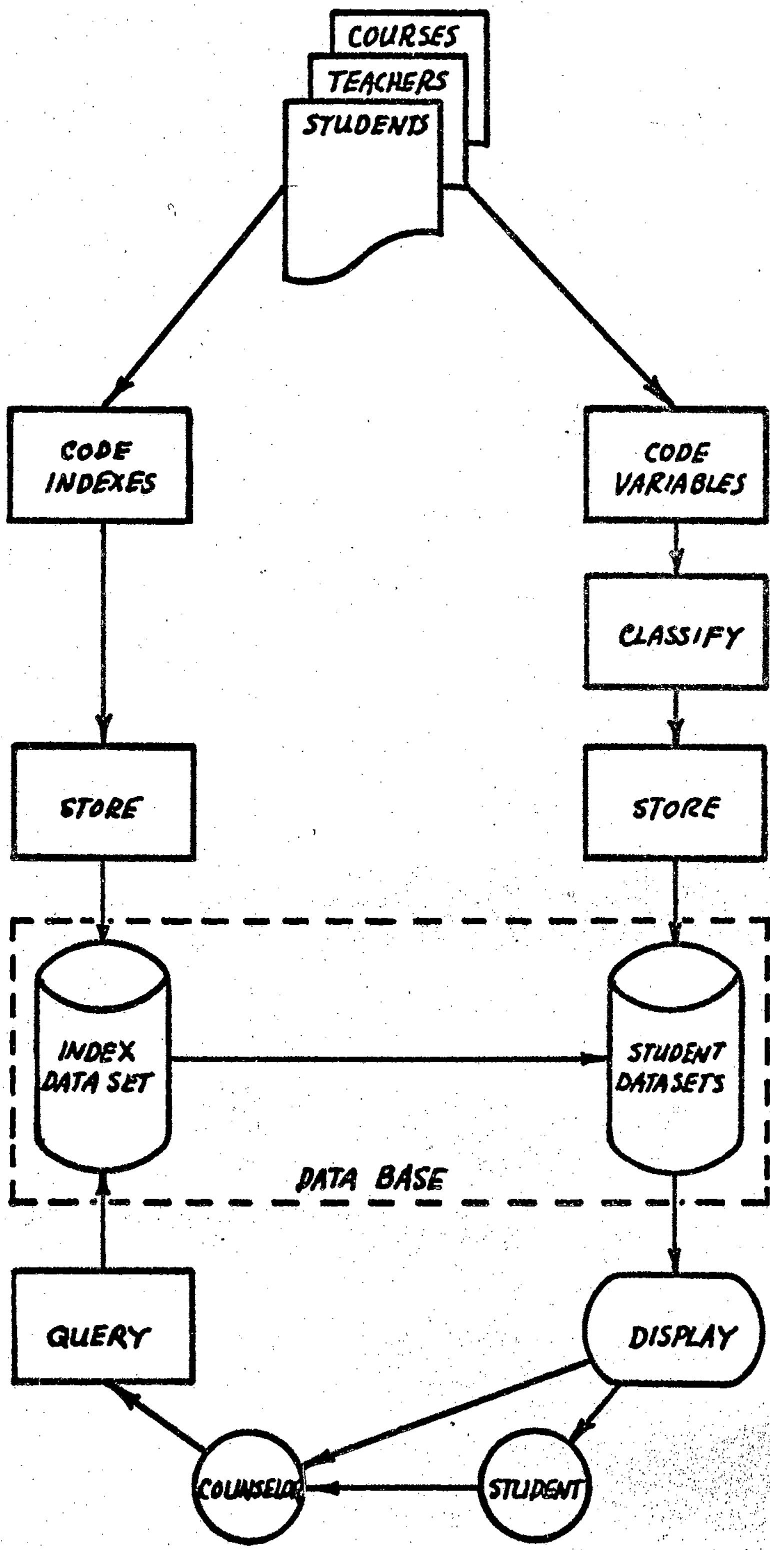


Figure 1. The Retrieval Process

which is simply a control field which uniquely identifies it for processing purposes. Concurrent with classifying the data, these indexes can be determined. The index data set and the student data set(s) may or may not be separate entities. Next is the problem of mapping the data onto a physical unit. Up to this point the data have been considered as "logical entities," i.e., "...information-carrying items and the relationships to one another" (37:124). Mapping refers to"...allocating elements of logical structure (items) to physical structure (cells), and managing the use of free space within the cell" (37:124). This step answers the question of file organization and is related to the storage medium and certain user requirements.

All of the selected student data are stored in the memory of the digital computer system. Once this has been accomplished, the counselor and the student can query the data base during the course of the counseling process. This is accomplished with a query language which permits the user to request selected data records. This language is one of the prime considerations of the data base design since it may place certain constraints on the selection of the file organization if a prewritten one is used.

One aspect of the information retrieval process which is not depicted in Figure 1 is that of file maintenance. File maintenance is an auxilliary set of operations needed to make changes to the data base. These operations may include the addition, deletion, modification, or rearrangement of data fields and/or records within a data set. More on this later, but file maintenance also can influence the design.

Record Structure

Basic to the concepts of the organization of information is that data are hierarchial, that is, symbols and words at a lower level can be combined to form a more generic classification. The letters of the alphabet are structured to form the student's last name and the ten decimal digits can be combined to form the student's identification number, for example. The set of rules used to form the combination is called syntax (36).

It was mentioned earlier that a record is a group of related data items which in turn are made up of data groups and elements. Meadows (11:153) explains that

> A field of a record is definable several ways and may be said not to be completely defined unless all these facets are specified.
>
> 1. Its position in the record, for example, characters 11-16 of a 100-character record
>
> 2. Its function within the record, such as a special end-of-record symbol, or tag
>
> 3. Its name, which is actually the name of the class of referents of the field such as a set of birthdates
>
> 4. Its value, or an individual referent or instance of the named referent set, as July 4, 1965

Therefore, it is possible to have field of decimal digit symbols located in positions 1-6 of the record, whose name is ID number, its function is to alphabetize the data set on student names, and whose value is 123456.

The remaining factors of record structure are concerned with storage and processing efficiencies--space and time aspects, so to speak. It is probable that several of the data elements within the record can vary in length, sometimes in great amounts. Take course titles as an example. Among a list of courses might be found titles like "Drama" and "Contemporary World Literature." The latter is a string of twenty nine symbols (including blanks) while the former has five symbols; a six-to-one ratio. If a fixed-length field were to be used for the course title, it would have to accommodate the longest title which probably does not occur very often. Thus, there can be a considerable waste of storage space. If memory space is at a premium, this problem can be reduced with at least three methods: (1) Use a variable length field, (2) abbreviate the data contents, or (3) assign an n-digit code to the data.

The use of variable length fields imposes a problem of locating the variable boundary. This can be solved by either preceding the variable length field with a count of the number of characters in the string or by placing a special character at the beginning or end of the string. In either case, there can be considerable storage savings at the expense of some increased processing time needed for the additional logical manipulations.

Abbreviation and coding amount to pretty much the same thing since they both involve the replacing of one symbol string with a shorter string. If abbreviations are used, precautions must be taken to guard against "semantic ambiguity," that is, "the plurality of meaning" of the abbreviation (36:158). An example of coding would be to use a three digit decimal number as a sequence code for the course title. This would provide one thousand unique codes if "000" is included. There are many other codes which can be used for this purpose but will not be discussed here. An excellent discussion on encoding can be found in Laden and Gildersleeve (31:187-204).

Record structure, then, involves the combining of symbol strings, called data items, in some logical way so that each has the same logical structure. Four facets are necessary to completely define an element; location, function, name, and value. Without all of these, meaningful information cannot be retrieved from the data base. The length of the element may be either fixed and/or variable making the entire record fixed or variable. Logically, fixed length records are easier to work with but at the expense of less efficient storage utilization. If all of the records are collected as a set, the grouping is called a data set or a file. File organization will be discussed next.

Data Base Structure

In order to completely define a file structure, several things must be known to the system:

1. The structure of the records in the file,

2. The logical sequence of records and their mapping onto the physical file,

3. The characteristics of the storage media,

4. The volatility of the records in the file, that is, the frequency of additions and deletions of records,

5. The activity or the frequency of referral to its records,

6. The size of the file, i.e., can the whole file be available to system at one time.

7. The growth potential of the file (5,28,36).

The main objective of data base organization is to systematically store data so that it can be retrieved in the fastest way possible at some later date (5). What is "fast" for one application may be "slow" for another. Therefore, several organizational methods exist because of the various types of storage media, computers, processing characteristics and file updates. "Different data organizations are therefore used to bridge the gap between the logical user requirements and the physical realities of storage media and computer hardware" (15:117). While attempting to bridge this gap, emphasis must be placed on both the function of the data base and the form or data base structure. It must be determined how accessible each piece of data is to be and how it should be placed in relation to other data in the file (30).

Processing

The function of the file system of this study is to provide a "data base" for the retrieval of student fact data. The types of operations which might be performed within the retrieval function are to (1) gain access to a student record, (2) insert a new student record, (3) delete a student record, (4) make changes to a record. Operation 1 will be called search and operations 2, 3, and 4 will be called update. The way in which these operations are performed defines the processing characteristics. The entity which initiates any of the above operations is called a transaction. Transactions are input to the system and have parameters relating to response time, sequence, and arrival rate.

The response time defines the maximum amount of time that the inquirer or the system can wait for the results of the desired operation. The search operation for this study requires a short response time, a matter of seconds, since the student and counselor will use the data in "real time." The update response time will be such that the transactions can be batched and processed daily, weekly, monthly, etc.

Sequence refers to the order in which the records can be placed prior to an operation. For the update operations it may be possible to place the transactions in the same logical order as the records in the data sets. If so, it is called sequential input. The search operation will no doubt process transactions as they occur since the desired response time will be a matter of seconds. The search transactions will no doubt occur at random as most real-life transactions do. This type of input is called random input.

The arrival rate defines the volume of transaction flow as a function of time. Arrival rate affects such things as batch size, computer size, and response time.

Storage Media

The "physical realities of storage media" play a definite role in how the data base is organized. The memories of today's general-purpose digital computer usually have three levels of accessibility; (1) equal access, (2) direct (random) access, and (3) sequential access (36).

Equal access is the highest level and it means that the contents of one addressable location can be retrieved in the same amount of time as any other location. Examples of this type of memory medium are the magnetic core, thin film, magnetic rod, and magnetic rope. This type of memory is considered the internal memory of the computer and is not ordinarily used for file storage due to its high cost.

Direct access means that to gain access to the next piece of data to be processed does not depend on the location from which the previous data were obtained. Direct access also means that there is some predictable relationship between the record key (student ID number, for example) and the physical storage address in the direct access device. Both of these meanings also apply to the equal access method but the basic difference is that the direct access method may require unwanted records to be read before the desired record is found while equal access can go straight to the location. Therefore, the access time of the direct access method is somewhat variable while that of the equal access method is constant. The direct access method is also called random access because it is highly suitable for processing transactions which arrive randomly and because any location chosen at random can be accessed equally as well as any other location. Examples of the direct access memory are magnetic disks, magnetic drums, and magnetic cards.

Sequential access means that in order to retrieve a record from a file, all of the records which physically precede it must be read first. Examples of sequential access memory media are magnetic tape, punched paper tape, and punched cards. For each new search sequence, the file must be physically "backed up" (except for tape systems with the read backward feature) to the starting point of the file prior to a sequential search through the file. It is not possible to jump to predetermined locations as it is with the equal and direct access media.

Organization Methods

The access characteristics of the storage media and the processing characteristics of the user's application suggests two basic types of data set organization, i.e., sequential organization and random organization.

Sequential organization. The simplest and most natural way to arrange the records in a file is to place them into a list one after another in contiguous storage locations. Only the physical location of the first record is known and the location of a record beyond the first can be found only after sequentially searching those which precede it. The search procedure may require that each record have a unique identifier called the key or index so that the "search logic" can match each key in the file with the desired key until the correct one is found. The records are usually, but not necessarily, stored in either ascending or descending sequence as determined by the value of the key. If the records are stored without keys, then they are placed in sequential storage addresses in the order in which they arrive into the system.

The main advantage of the sequential organization is that it provides an efficient method for processing if the transactions can be batched and placed in the same order as the data base file. When compared to the linked list below, sequential organization utilizes less storage because there is no need to set aside space for the link. On the negative side, updating a record in the file is difficult. To insert a new record, all records higher in the sequence must be "pushed down" to make room for the insertion and this requires that the entire file be copied. To delete a record, the file must be "pushed together," again by copying the entire file. The search and update operations become very inefficient if the transaction activity is low and the record volatility is high.

All forms of storage media can accommodate the sequential organization but some of the devices can take advantage of more sophisticated search techniques applicable to a sequential list. For example, all direct access media will allow the use of the binary search, a procedure whereby the record is located by successively halving the search interval until the record is found (the file must be ordered on the search key).

Random organization. This type of organization characterizes a procedure where the value of the record key is used to calculate the storage address. As mentioned earlier in this chapter, direct access implies that there is some predictable relationship between the record key and the physical storage address. The record key can be used directly or it can be used as input to a transformation algorithm which serves to compress the key into a more reasonable address range. These two methods will be termed direct and indirect file addressing (28).

A directly addressed file includes the use of the key as the address in a one-to-one correspondence or the use of a cross-reference table. The cross-reference method requires a table containing the key and an assigned address. The table may be another data set internal to the system or it may be an external printed one. In either case, a table look-up is required. Strict controls must be maintained to keep the list accurate and up-to-date. Using the key as the address requires that the key be numeric and that the record be fixed length. This method reserves a fixed length location of storage for every key within its range. For example, if the key ranged from 00 to 99 inclusive then one hundred locations are set aside. This method has the disadvantage of wasteful storage if all the keys are not used. A six digit student ID number with a range of 000001 through 500000 would require 500,000 locations in the file. Obviously this method cannot be used economically for 4,000 students.

An alternative technique which can be used to compress the large original range of keys into a smaller range of addresses by computing an indirect address. This procedure is called randomizing. Randomizing has become a very common method of addressing. It locates the majority of records but never all of them with one file reference. It is quicker than the file scanning or table look-up methods. One disadvantage is that the randomizing method may not yield a high file density. Seventy to eighty per cent is a reasonable figure to attain the cause of the "low density" arises as a result of the synonym problem, discussed below, which can be alleviated by allocating more storage than needed. Another disadvantage is that it will not necessarily produce a unique address, i.e., synonyms will probably occur. Synonyms are record keys which randomize (transform) to the same file address. There are several randomizing algorithms which can be used. Some of them are:

1. Truncation: The quickest form of truncation is to divide by 100, 1,000, 10,000, etc. This is effectively truncating the reference number and retaining the required number of low order digits. If the set of reference numbers is irregular this method will produce many overflows.

2. Folding: Truncation throws away some of the information contained in the key set and any information discarded tends to weaken the uniqueness of the key set and to weaken its distribution. A better method would be to add parts of the reference number to each other to produce the desired address.

3. Division: If the range of addresses is N, that is N address locations are available with consecutive numbering, a reference number

may be divided by N and the remainder used as the randomized number. If P is the lowest address, this is added to the remainder.

4. Division by a Prime Number: With most sets of keys the above division method gives better results if the division is by a prime number. This prime number should be slightly less than the number of available storage locations. This method does not always give the best results, but it usually produces fewer overflows.

5. Radix Conversion: The radix of a number may be converted, to radix II for example, and the excess high order digits truncated. This method is superior to simple truncation.

6. Squaring: The number may be squared and the center digits of the result used. The disadvantage of this method is it can produce an excessive number of zeros.

It cannot be stated in advance which randomizing method will work best. Each must be evaluated for the specific application. In doing so, the following statistics must be brought into balance: (1) "every possible key in the file must randomize to an address in the allotted range, and (2) the addresses should be distributed evenly across the range so that there are few synonyms" (28:42). Figure 2 schematically describes the randomizing technique.

The main advantage of the random organization is that a record can be retrieved from the file without retrieving other records first. However, it is not suited for retrieving large batches of records as would be required for the update operations. Also there are overhead and programming difficulties associated with synonym overflow in the case of indirect addressing method and the scanning of large tables which may be created in the case of direct addressing method (15).

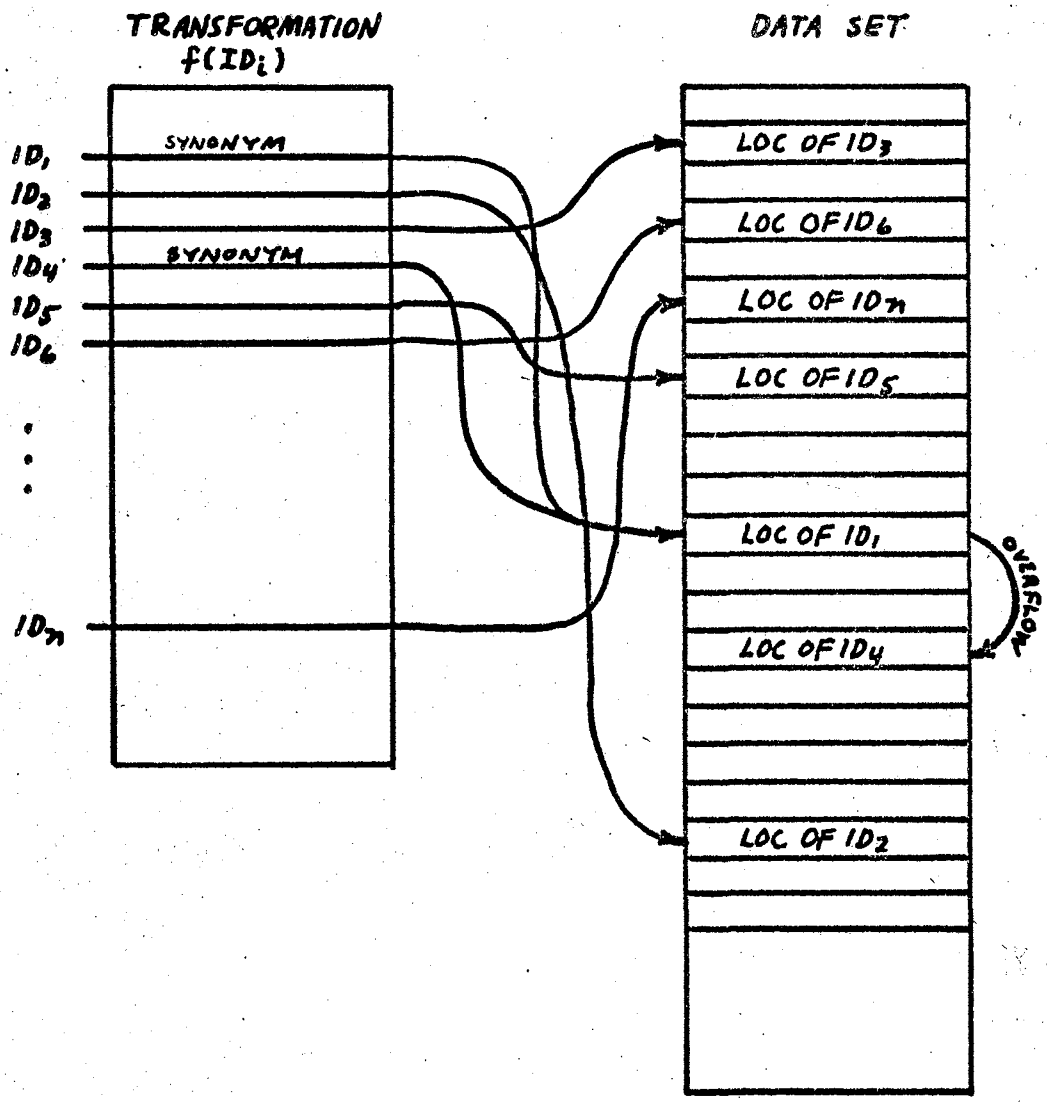


Figure 2. Schematic of the Randomizing Technique

Linked List. To overcome the problems of inserting and deleting records of the sequential organization, the linked list organization was derived. This method requires that each record carry a link which is used to locate the next record in the list. A link is an address, pointer, index, or reference to another record. The linked list method is more flexible since the successive items in the file need not be placed in contiguous locations--any arbitrarily chosen available location can be linked to its predecessor. An immediate disadvantage of the linked list is the additional storage space required for the links. In some situations this might be an important factor but in other situations there might be less storage used if the linked list can reduce the overlapping of table data by sharing the common segments. Insertion and deletion of records in the linked list is much easier and more efficient since the files do not have to be copied. For example, Figure 3 shows how only two links need to be changed in order to insert record 15 between records 14 and 17. The link of record 14 would be changed to point to the location of record 15 and the link of the new record 15 would be set to point to record 17. Deletion of record 15 would involve only one change, viz., change the link of record 14 to point to record 17. These changes in the linked list require considerable less effort than that of the sequential list. Another disadvantage os the linked list is that random access may not be possible. Recall that for random access there is a predictable relationship between the record key and the file address. "The basic concept of a list is that pointers are used to divorce the logical organization from the physical organization" (15:122). The usual criterion for selecting a location for insertion of a record is to pick a location from a pool of available locations. The "randomized" address may not be in the pool. It should be noted, however, the link concept can be used to handle the synonym overflow problem created by the randomizing technique (see references 15 and 28).

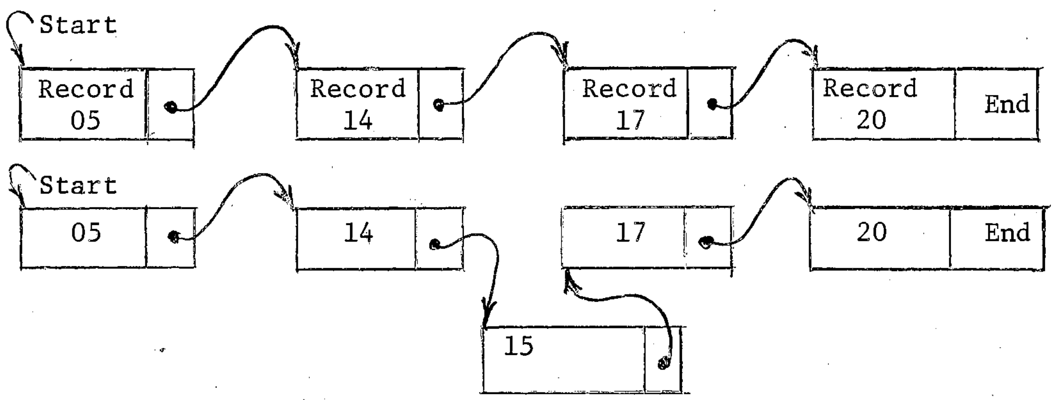


Figure 3. Insertion of a Record; Linked List

Branching File Structure. Bachman and Williams (7) have suggested an organization called branching file structure which is an extension of the linked list concept. The purpose of this organization is to save storage by removing "often-recurring values" from a main file and replacing them with much shorter links or pointers to entries in an auxiliary file. Consider a student data file which contains a list of courses that represent his current enrollment schedule. Figure 4a shows how a conventional file might look with its repeated course data. Figure 4b shows (1) how the course names are removed and put into an auxiliary file and (2) how the storage requirements of the main file are reduced by course name field size times the number of courses

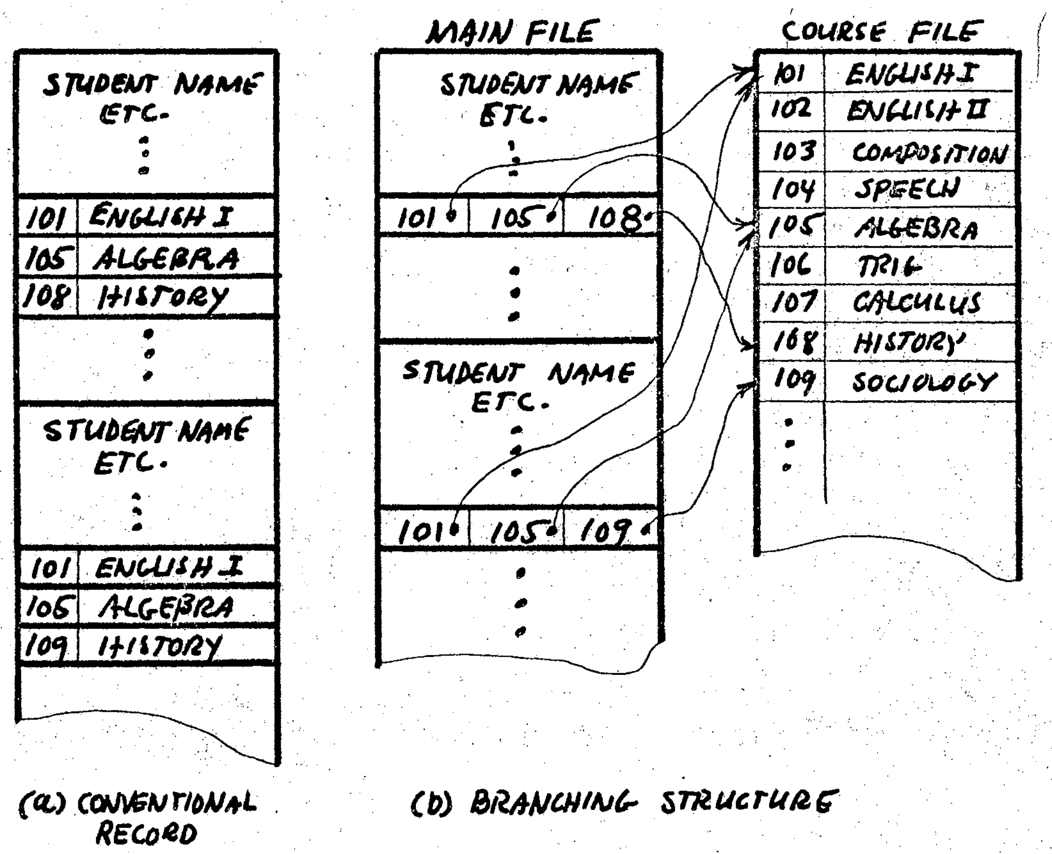


Figure 4. Branching File Structure

stored there. Of course, these savings do not come absolutely free since the Course File will require a certain amount of storage and there will need to be additional programming logic to access the auxiliary file.

Indexed Sequential. The indexed sequential file organization is a sequentially recorded data set with pseudo random attributes. This method has popular appeal because:

1. Its sequential organization results in more efficient sequential processing.

2. Its sequential organization provides more efficient searches for logical groups of records (for example, all students whose ID number begins with 543).

3. It has the appearance of a randomly organized data set.

When the indexed sequential file is created, the records are first sorted on their key and placed on the direct-access device in sequential order. As the file is created an index of "certain critical reference points" is also created and stored in the data set itself or one separate from the information data set. The index is a multilevel list which serves to localize the record as to volume, cylinder, and region (for disks). Retrieval without scanning the entire file is accomplished by searching the index until the correct region is found. The region is then scanned sequentially until the desired record is found. This method is a variation of the directly addressed file utilizing the table look-up approach mentioned earlier.

To overcome the "recopy" problem of the sequential file, the indexed sequential method divides the information data set into two parts; a prime area and the overflow area. When the file is created the records are written in the prime area in sequential order. Later when a record is to be added to the file an attempt is made to put it into the prime area if its respective storage location is not already occupied. If a location is not available in the prime area, space is made available by moving a record from the prime area to the overflow area. All of the overflow records are linked together with the final record in the chain pointing back to the index and, therefore, the logical sequence of the file is maintained. When deletions are made, the record is replaced with a "dummy" which cannot be reused. During the course of using the file several additions and deletions will probably be made causing many "holes" in the prime area and the overflow area to fill up. Therefore, the file should be reorganized periodically. Both additions and deletions tend to increase the search time.

The above list of file organizations should not be considered as conclusive. Since the selection of a file organization is so highly application oriented, the references made in this section apply to factors which are considered applicable to this study. A summary of some of these factors are listed below.

1. The response time requirements of the various types of inputs.

2. The type of search key and other data available for retrieval and updating.

3. The link or chaining requirements between subsets of data.

4. Data compactness versus available storage space. It may not be worth the extra cost of "packing" the data if sufficient storage is available (39).

5. The number of indices and amount of data redundancy as determined from priority, response requirements, and frequency of response

at the terminal.

6. The configuration and speeds of the computer which affect file organization and response time (16).

Language Considerations

The term language as used in this section refers to the set of symbols, words, and phrases which are used to communicate with the computer systems as opposed to a "natural language" such as English. This language will be called, for the purpose of this study, the man-machine language. The man-machine language must serve two functions in this retrieval system; one is file maintenance and the other is query. File maintenance function includes file creation and file update (additions, deletions, and modifications). The query function will permit the "...user to request selective data records from the file, extract data from the records, and sort and format the output" (37:143). Both functions may be implemented in a single man-machine language or they may be implemented separately. For example, an assembler-level language may be used for both functions or PL/I could be used for file management and FASTER* used for query.

Unfortunately there is no single language development to date which is best suited for all retrieval operations and file organizations. The one that comes the closest is PL/I because it provides for (1) string manipulation (bits and characters), (2) list manipulation (linked lists and pointers), (3) the handling of sequential, random, and indexed sequential file organizations, (4) recursive programming,

---

*FASTER is a Type 3 IBM man-machine language designed to provide filing and source data entry techniques for easier retrieval.

and (5) the usual logical, algebraic, and matrix operations. One important facet not included in PL/I is an interface between the query programs and the teleprocessing facility. These operations must be programmed in assembler language and be called by PL/I as subroutines.

An example of a language which does provide a teleprocessing interface is the FASTER system (16). The language was written to automatically perform the basic teleprocessing functions required for remote operations. FASTER supports the indexed sequential method of organization and fixed length records only.

The man-machine language must focus on all the aspects of a retrieval system. Therefore, "only a study of the language facilities, the type of hardware to be used, the types of input queries and the way they will be submitted to the system, the capabilities of the storage media, and the computer efficiencies will dictate which language to use for programming a particular data management system" (15:131).

CHAPTER IV

DATA BASE DESIGN

Introduction

This chapter describes the approach used to define the data base model and answers several of the questions posed in Chapter I. The approach was to identify the student variables to be stored in terms of the user's (student, counselor, teacher, and administrator) needs, design the data items to be stored, identify the data sets, determine the record structure and arrangement, and determine the arrangement of the files or data sets within the data base.

All of the above facets of the data base design were primarily related to the choice of the man-machine language to be used for query. This language is the FASTER programming system mentioned briefly in Chapter III. More specifically,

> FASTER is a programming system for retrieval and maintenance of indexed sequential files from remote terminals. The methods are applicable to a wide variety of applications. FASTER incorporates a macro-language which provides for the writing of message processing programs on a functional level.
>
> Some of the major functions supported include: retrieval of records from indexed sequential files, modifications and additions of ISAM records, data manipulation (including Boolean logic capability), formatting of responses to the selected terminals..., message switching with fixed dynamic routing, and the recording of audit data on a system logging device (16:33).

The FASTER language places the following constraints on the design of a data base:

1. ISAM (indexed sequential access method) organization must be used.

2. Fixed length records must be employed.

3. A direct access storage device must be available.

FASTER was selected for project TGISS because it is supported by IBM for the machine configuration described earlier and it eliminates the need for involvement in the nitty-gritty of telecommunications line and message control programming. The latter aspect resulted in significantly less programming effort which was required to put the application "on the air." This reduction was offset by the disadvantages of the indexed sequential method of the file organization and the relative inefficiency of the coding produced by FASTER, but not sufficient in magnitude to be of concern.

Identifying the Variables

The data base design began with the requirements of the user. The data that were to be stored was selected and organized in a manner which was to satisfy his retrieval demands. The contents of the data base for the information retrieval aspect of TGISS were the result of two work conferences held in Bartlesville and Stillwater, Oklahoma. The first work conference included personnel from project TGISS, counselors and administrators from the Bartlesville schools, representtatives from surrounding public schools, and three outside consultants. The purpose of the first session was to determine the general user requirements and to identify the variables to be included in the data base. Some of the important results of this meeting pointed out the data needs and the benefits which would result from a computerized retrieval system. These were:

1. Access to data such as health, attendance, grade point averages, etc., traditionally outside the counselor's office tends to promote (a) increased interaction between the counseling office and auxilliary services in the school, (b) leads to restructuring the school organization consistent with the systems approach, and (c) develops a better communication system between the counseling office, the administration, and auxiliary services.

2. The counselor should periodically cycle through the data base searching for trends in student mobility with emphasis on possible discriminatory practices. Are students equally mobile academically with respect to race, socio-economic status, ethnic background, etc.?

3. Periodic cycling the data base would highlight areas for curriculum reform, e.g., students with low scores on achievement tests suggests a need for remedial education programs so that student graduation is not delayed.

4. There should be a capability to call information, one item or a group of items, on an individual student or a group of students. Such information would be used to produce (a) nine-week progress reports, (b) a midterm and end-of-semester list of failures to allow the counselor to act before the fact in order to prevent the student from being "pushed" out of the system, (c) a list of students who do not meet graduation requirements, (d) detectors for redirecting the student to another major so he can graduate, (e) a list of students receiving failing grades in required courses, (f) a list of students who have "excessive" absences and tardiness, (g) a list of potential dropouts, (h) an indication of a change in the performance status of the student,

and (i) a list of students who must meet particular deadlines, e.g., applications for college.

5. The data base should include a record of the counselor's contact with the retrieval system to provide feedback of usage patterns to be used for possible revisions to the system.

6. The contents of the data base should be secure against unauthorized entry.

7. Successive data base generations should permit additions or deletions of items and files. The initial design is predicated on the counselor's present conception of counseling in relation to the computer system. The initial prototype will modify present system requirements in such a manner that the counselors and other school personnel may want information not specified in the initial design effort.

During the course of the work conference discussions and from the above list of needs, a tentative list of student variables were identified and tabulated for further scrutiny. This was done at a second work conference. The participants were the project (TGISS) personnel, a selected group of Bartlesville counselors, and one outside consultant. The approach taken at this meeting was to determine the data aggregates which the counselor would like to see displayed on the graphic display devices (CRT and typewriter) to be supplied with the system. Sixteen data display groupings were identified as a result:

1. Transcript Display
2. GPA (honor point averages) Display
3. Preconference Data List
4. Current Student Schedule
5. New Student Schedule

6. Absence Profile
7. Grade (GPA) Trends
8. Standardized Test Profile
9. Interests/Preferences/Honors
10. Work Experience Display
11. Address Labels
12. Unsatisfactory Grade Reports
13. Ten Item Health Diagnostic Profile
14. Subject Area Achievement vs. Aptitude
15. Health Record
16. Reports and Summaries of Selected Data

In addition, the following 67 variables were identified to implement the displays.

General

1. Student identification number
2. Student's name
3. Sex
4. Race
5. Birthdate
6. Home address
7. Home phone
8. Grade level, current
9. High school track
10. School
11. Religious preference ?
12. School transferred from
13. Year graduated from 6th grade

14. Year graduated from 8th grade

15. Father's name

16. Father's employer

17. Father's business phone

18. Mother's name

19. Mother's employer

20. Mother's business phone

21. Family doctor's name

22. Family doctor's phone

Transcript (7th and 8th grades)

23. Course number

24. Course name

25. Honor points - 7th grade

26. Honor points - 8th grade

Transcript (grades 9, 10, 11, and 12)

27. Class (9th, 10th, 11th, or 12th)

28. Course number

29. Course name

30. Honor points, first semester

31. Honor points, second semester

32. Earned credits

Absences

33. Cumulative half-days

34. Half-days by week

Grade point averages

35. Junior High

36. Senior High

37. Academic Area

38. Grade 7

39. Grade 8

40. Semi-semester (starting with the ninth grade)

Course Schedule

41. Period

42. Course number

43. Course name

44. Location

45. Teacher

Health Profile

46. Eyes

47. Ears

48. Teeth

49. Growth

50. Heart

51. Respiration

52. Blood

53. Skin

54. Weight

55. Feet

Health Record

56. Name of disease

57. Date of disease

58. Vaccination date

Standardized Test Scores

59. Kuhlman-Anderson

60. SRA

61. DAT

Work Experience

62. Job title

63. Experience

64. Status

Interests, Preferences or Honors

65. Description

66. Type code

Preconference Data

67. Description (any data entered by the counselor)

Data Item Design

The design of the data items consisted of determining (1) what fields to include in the item, (2) the arrangement of these fields within the item, (3) the size of the field, and (4) the mode of the contents.

Two criteria were used in the item design. One was based on the language to be used for retrieval of the stored data. The FASTER system requires the use of the indexed sequential file organization and fixed length records. The other was storage space which was reduced by coding or abbreviating certain data item elements and the use of a branching file structure.

Initial Data Item Formats

The data items of the student data base were arbitrarily named as follows.

1. Demographic Data
2. Transcript I (grades 7 and 8)
3. Transcript II (grades 9 through 12)
4. Absence Trends
5. Grade Trends
6. Current Schedule
7. Health Profile
8. Health Record
9. Standard Tests
10. Work Experience
11. IPH Data (interests, preferences, honors)
12. Privileged Data

The data groups and elements for each of the above data items are shown in Tables I through XII.

Each table contains the name of the element; the codes which have been assigned to certain attribute data; the length of the element in characters; and the mode, where "A" means alphabetic only (letters A through Z and any special character including the "blank"), "N" means digits only (0 through 9), "AN" means alphabetic/numeric. The length of each element was determined from the source documents obtained from the public schools in Bartlesville. Since the record length had to be fixed, so did the element length. Therefore, the length of the longest source item was selected as the element length. Provisions were made to reduce this length by abbreviation and coding where appropriate. The indication of abbreviation is shown in the code column using the symbol ABBR.

Discussion of the Data Items

1. Demographic Data, Table I. Strictly speaking, demographic data refers to those data which are essentially sociological such as "age, sex, socio-economic status of parents," etc. (10:20). The term is used here somewhat loosely but, since the contents of this data item is mostly demographic, the name seems appropriate.

The criterion for placing elements of data into this item was to simply include those which could not be logically placed elsewhere.

The student identification number (ID) was used to uniquely identify the student and to serve as the primary search key. The ID had already been assigned to the student by the school system. The elements sex, race, birthdate, high school track, and school were coded in order to save space. This space saving was offset by an increase in processing time due to the "table look-ups" needed to translate the code to a more meaningful form. This increase was slight, however, since most codes were used directly due to either the conventional use of the code, such as 032069 for March 20, 1969, or counselor training.

The total length of the Demographic Data item is 212 characters.

2. Transcript I, Table II. Transcript I contains the course numbers and descriptions, and the earned grades (in honor points) for the 7th and 8th grades. This data item contained fifteen repeating data groups of seventeen characters each for a total of 255.

3. Transcript II, Table III. This data item is similar to Transcript I but in addition it contained the grade explicitly (in Transcript I the grade was implied by its location), honor points by semester, and the earned credits. An analysis of the existing transcripts revealed that twenty-four repeating data groups were required

TABLE I

ELEMENTS OF THE DEMOGRAPHIC DATA ITEM

Element Name	Code	Length(Char)	Mode
1. Delete*	--	1	N
2. Student identification number	--	6	N
3. Student name (last, first, middle)	--	23	A
4. Sex	Male=M Female=F	1	A
5. Race	Caucasian =C Indian =I Negro =N Spanish American=S Other =0	1	A
6. Birth month	2 digit code; 01-12	2	N
Birth day	2 digit code; 01-31	2	N
Birth year	last 2 digits of year	2	N
7. Home address			
-house number	--	4	N
-direction	ABBR	2	AN
-street	ABBR	11	AN
8. Home phone	--	7	N
9. Current grade	--	2	N
10. High school track	College =C Vocational=V General =G Special =S	1	A
11. School	College High=C Sooner High =S	1	A
12. Religious Preference	ABBR	15	A
13. School transferred from	ABBR	15	A
14. Year graduated: 6th	last 2 digits of year	2	N
15. Year graduated: 8th	last 2 digits of year	2	N

TABLE I (Continued)

Element Name	Code	Length(Char)	Mode
16. Father's name	--	20	A
17. Father's employer	ABBR	15	AN
18. Father's business phone (station, extension)	--	11	N
19. Mother's name	--	20	A
20. Mother's employer	ABBR	15	AN
21. Mother's business phone (station, extension)	--	11	N
22. Family doctor's last name	--	13	A
23. Family doctor's phone	--	7	N

*The delete character is required by the FASTER maintenance function.

TABLE II

ELEMENTS OF THE TRANSCRIPT I DATA ITEM

Element Name	Code	Length(Char)	Mode
1. Course number	--	4	N
2. Course name	ABBR	11	AN
3. Honor points, grade 7	A = 4 B = 3 C = 2 D = 1 F = 0	1	N
4. Honor points, grade 8	(same as element 3)	1	N

NOTE: The above data group occurs fifteen times.

TABLE III

ELEMENTS OF THE TRANSCRIPT II DATA ITEM

Element Name	Code	Length(Char)	Mode
1. Grade	9th grade = 0 10th grade = 1 11th grade = 2 12th grade = 3	1	N
2. Course number	--	4	N
3. Course name	ABBR	12	AN
4. Honor points, first semester	A = 4 B = 3 C = 2 D = 1 F = 0	1	N
5. Honor points, second semester	(same as element 4)	1	N
6. Earned credits	--	2	N

NOTE: The above data group occurs twenty-four times.

for this data item. Each data group contains twenty-one characters and a total of 504.

4. Absence Trends, Table IV. Absences are to be recorded in half-days on a weekly basis resulting in a number range of 0-10. The corresponding cumulative absences number range is 0-360. The weekly absences element repeats thirty-six times for the academic year. The resulting total for this data item is twenty-five.

5. Grade Trends, Table V. Grade trends are recorded as cumulative credits and honor points which can be used to compute the GPA. Starting with the ninth grade, the GPA data will be recorded every nine weeks, or sixteen times during grades nine through twelve. The academic area GPA data included only those courses from the students area of emphasis such as math, science, home economics, etc. The total character requirement is eighty-four.

6. Current or New Schedule, Table VI. The format of this data item is used for both the new (next semester's) or the current semester's course schedule. An analysis of the student's schedules revealed that a maximum of seven courses would be taken by some of the students, therefore seven repeating data groups are provided. The total number of characters for the item is 231.

7. Health Profile, Table VII. The Health Provile consists of ten characteristics measured along a three-point scale as shown in the table. The length of this data item is ten.

8. Health Record, Table VIII. This data item is used to record the diseases that the student might have had or vaccinated against. The disease code, two digits, allows for up to ninety-nine diseases although only ten are shown for illustration. An analysis of the

TABLE IV

ELEMENTS OF THE ABSENCE TRENDS DATA ITEM

Element Name	Code	Length(Char)	Mode
1. Cumulative absences	--	3	N
2. Weekly absences	--	2	N

NOTE: Element 2 occurs thirty-six times; once for each week of the semester.

TABLE V

ELEMENTS OF THE GRADE TRENDS DATA ITEM

Element Name	Code	Length(Char)	Mode
1. Grade 7	--	4 (Note 2)	N
2. Grade 8	--	4	N
3. Academic Area	--	4	N
4. Junior High Cumulative	--	4	N
5. Senior High Cumulative	--	4	N
6. Semi-semester (starting with the 9th grade) (Note 1)	--	4	N

NOTE 1: Element 6 occurs sixteen times.

NOTE 2: The first two positions contain the cumulative credits and the second two positions contain the cumulative honor points.

TABLE VI

ELEMENTS OF THE CURRENT OR NEW SCHEDULE DATA ITEM

Element Name	Code	Length(Char)	Mode
1. Period of the day	8:00-9:00 = 1 9:05-10:25= 2 10:30-11:30= 3 11:35-1:05 = 4 1:10-2:10 = 5 2:15-3:15 = 6 3:15-3:45 = 7	1	N
2. Course number	--	4	N
3. Course name	ABBR	12	AN
4. Location	--	3	AN
5. Teacher number	--	3	N
6. Teacher name	(last name only)	10	A

NOTE: The above data group occurs seven times.

TABLE VII

ELEMENTS OF THE HEALTH PROFILE DATA ITEM

Element Name	Code	Length(Char)	Mode
1. Eyes	Normal = 3 Subnormal = 2 Very poor = 1	1	N
2. Ears	*	1	N
3. Teeth	*	1	N
4. Growth	*	1	N
5. Heart	*	1	N
6. Respiration	*	1	N
7. Blood	*	1	N
8. Skin	*	1	N
9. Weight	*	1	N
10. Feet	*	1	N

*The codes for all elements are the same.

TABLE VIII

ELEMENTS OF THE HEALTH RECORD DATA ITEM

Element Name	Code	Length(Char)	Mode
1. Disease code	See code list below	2	N
2. Name of disease	ABBR	15	A
3. Date of disease	Numeric code for month (2 digits) and year (last 2 digits)	4	N
4. Date of vaccination	(same as element 3)	4	N

NOTE: The above data group occurs seven times.

Code List:

Chicken Pox = 1
Diphtheria = 2
Measles = 3
Mumps = 4
Scarlet Fever = 5
Smallpox = 6
Polio = 7
Asthma = 8
Whopping Cough = 9
Tetanus Toxoid = 10

student's cumulative records indicate that space for seven diseases should be allowed for. The total item length is 175.

9. Standard Tests, Table IX. Three standardized tests are being administered at the Bartlesville public schools as shown in the table. The length of this item is forty-nine.

10. Work Experience, Table X. The work experience data includes an abbreviated job description, length of experience in months, and a status code. If the job was voluntary or non-paying, it is given the code "V". If it was a paying job, that is, voluntary, it is given the code "N." Each data group is twenty-three characters in length, or a total of 138 for the data item.

11. IPH Data, Table XI. The IPH data item consists of (a) interests such as hobbies, extra-school activities, etc., (b) preferences pertinent to a vocation or other post-high school ambitions, and (c) honors, either academic or non-academic. Each entry was type coded as shown in the table. The data group is repeated five times yielding a total of 205 characters.

12. Privileged Data, Table XII. Up to 240 characters of data can be entered in free form by the counselor. This constitutes a "page" of storage with one allocated for each student assigned to the counselor. The set of "pages" belonging to the counselor make up his "book" of privileged data. All of the counselor's "books" are placed in a single data set, but the contents belonging to an individual counselor can be accessed only by him assuming he has made proper use of the "unlocking" algorithm. The counselor has the facility to record and erase all or part of the contents of his book at will. The length of each page is 270 characters.

TABLE IX

ELEMENTS OF THE STANDARD TESTS DATA ITEM

Element Name	Code	Length(Char)	Mode
1. Kuhlman-Anderson	--	3	N
2. SRA Part 1	--	2	N
3. SRA Part 2	--	2	N
4. SRA Part 3	--	2	N
5. SRA Part 4	--	2	N
6. SRA Part 5	--	2	N
7. SRA Part 6	--	2	N
8. SRA Part 7	--	2	N
9. SRA Part 8	--	2	N
10. SRA Part 9	--	2	N
11. SRA Part 10	--	2	N
12. SRA Part 11	--	2	N
13. SRA Part 12	--	2	N
14. SRA Part 13	--	2	N
15. SRA Composite	--	2	N
16. DAT Part 1	--	2	N
17. DAT Part 2	--	2	N
18. DAT Part 3	--	2	N
19. DAT Part 4	--	2	N
20. DAT Part 5	--	2	N
21. DAT Part 6	--	2	N
22. DAT Part 7	--	2	N
23. DAT Part 8	--	2	N
24. DAT Part 9	--	2	N

TABLE X

ELEMENTS OF THE WORK EXPERIENCE DATA ITEM

Element Name	Code	Length(Char)	Mode
1. Job description	ABBR	20	AN
2. Months of experience	--	2	N
3. Status	Non-paying = V Paid = N	1	A

NOTE: The above data group occurs six times.

TABLE XI

ELEMENTS OF THE IPH DATA ITEM
(Interests, Preferences, Honors)

Element Name	Code	Length(Char)	Mode
1. Description	ABBR	40	AN
2. Type Code	Interest = I Preference = P Honor = H	1	A

NOTE: The above data group occurs five times.

TABLE XII

ELEMENTS OF THE PRIVILEGED DATA ITEM

Element Name	Code	Length(Char)	Mode
1. Delete	--	1	AN
2. Counselor ID	--	2	N
3. Student ID	--	6	N
4. Overflow digit	--	1	N
5. One line of descriptive data as entered by the counselor	ABBR	80	AN

NOTE: The above data item occurs three times.

Data Item Summary

Table XIII recapitulates the size of the data items. This table shows the sizes of the data items as they appeared at this particular point of the design of the data base but not necessarily the size in the final data set records. This first step was performed to determine what the student record would contain if it were organized as a two-file data base; one file containing the "student" data of Tables I-XI and the other file containing the counselor's privileged data. With this simple file arrangement, the records of the student file would contain 1938 characters each and the counselor file record would be 270 characters in length.

TABLE XIII

RECAP OF DATA ITEM SIZES

Part	Data Item Name	Size (Characters)	
1.	Demographic Data	212	
2.	Transcript I	255	
3.	Transcript II	504	
4.	Absence Trends	75	
5.	Grade Trends	84	
6.	Current or New Schedule	231	
7.	Health Profile	10	
8.	Health Record	175	
9.	Standard Tests	49	
10.	Work Experience	138	
11.	IPH DATA	205	1,938
12.	Privileged Data	270	270
TOTAL			2,208

Data Base Organization

With the elements of the data items firmly established, the next step was to determine how the data elements should be organized within the data base. The questions to be resolved were:

1. How should the data records be structured?
2. How should the data records be arranged within the data file?
3. How many data files should be included in the data base?
4. How should the data files be arranged within the data base?

Having already established that (1) the file access method is to be indexed sequential, (2) the records must be fixed length and (3) the storage medium is of the disk variety, it remained to answer the above questions in terms of storage utilization and the constraints imposed by the indexed method.

Storage Space Reduction

The simplest data base organization which could be developed from the previously defined data items would be one consisting of the two data files discussed in the previous section. However, a close examination of the data items revealed an excessive amount of redundancy of data values for those elements which were not unique to the individual. For example, it was obvious the same course name would appear many times throughout the file in the transcripts and course schedule. The same was true for religious preference, family doctor, teacher's name, and diseases. Therefore, it was proposed to remove these elements from the data item and replace them with a linkage to an entry into another file. In some cases the linkage was already present and it was only necessary to remove the element. The course number is an example.

The result was a considerable savings of storage. Table XIV shows the data item, element name and number, previous storage requirement of the element, the number of positions to be added for the link, the net reduction after compensating for the link, the repetition factor, i.e., the number of times that the element is repeated in the data group, the gross storage reduction for the repeated element, and the total gross storage reduction for each student. Note the reduction of 743 characters making the student record 1195 characters in length. Assuming a student population of 4,215, this meant a gross reduction of 3,131,745 character positions.

Data Sets

This gross savings of storage was offset by the storage requirements of the table data sets which resulted from the removal of the elements. Table XV exhibits the resulting table data sets including the data set name, field or element name, element length, record length, estimated repetition factor, data set size, and total storage requirements. The repetition factor was considered an estimate due to the dynamic nature of the number of teachers, doctors, etc. After deducting the storage required by the table data sets from the gross reduction of the student data set, the net savings are over three million data characters, or about twenty-one cylinders.

The proposed data base now consists of the following data sets:

1. Student File
2. Course File

TABLE XIV

DATA ELEMENTS WHICH WERE REMOVED FROM THE STUDENT RECORD

Data Item Name	Element Name	Previous Storage Requirement	Linkage Requirement	Net Decrease	Repetition Factor	Gross Reduction
Demographic Data	Religious Preference (#12)	15	2	13	1	13
	Doctor and Phone (#22 and #23)	20	2	18	1	18
Transcript I	Course name (#2)	11	0	11	15	165
Transcript II	Course name (#3)	12	0	12	24	288
Current Schedule*	Course name (#3)	12	0	12	7	84
	Teacher's name (#6)	10	0	10	7	70
Health Record	Name of Disease (#2)	15	0	15	7	105
Gross decrease per student						743

*Only the Current Schedule was considered here since the New Schedule was recommended as a separate "transient" data set which exists during the period between enrollment and the next semester.

TABLE XV

TABLE DATA SETS

Data Set Name	Element Name	Element Length	Record Length	Estimated Repetition Factor	Estimated Data Set Size
Course	1. Delete*	1	17	223	3,791
	2. Course number	4			
	3. Course name	12			
Teacher	1. Delete*	1	14	238	3,332
	2. Teacher ID number	3			
	3. Teacher name	10			
Doctor	1. Delete*	1	23	28	644
	2. Doctor ID number	2			
	3. Doctor's name	13			
	4. Phone number	7			
Disease	1. Delete*	1	18	14	252
	2. Disease ID	2			
	3. Disease name				
Church	1. Delete*	1	18	26	468
	2. Religion ID	2			
	3. Religion name	15			
	Total				8,487

* The delete character is required by the FASTER maintenance function.

3. Teacher File
4. Doctor File
5. Disease File
6. Church File
7. Privileged Data File

Record Structures and Final Data Item Design

The arrangement of the data items within the record is abritrary except for the "delete" character which must be in the first character position, and the search key which must be in contiguous positions. A search key must exist to take advantage of the direct access facility of the indexed sequential access method; without it ISAM would sequentially search the entire file until the desired record is found. Tables XVI-XXII describe the records of each of the data sets in terms of the four "facets" discussed in Chapter III, viz., name function, position, and value. The function and contents of the field or element is implied by its name unless otherwise specified. The repeating data groups are defined in detail for the first group only; the remaining groups of the data item are lumped together and shown in contiguous positions. Due to the length and complexity of the student file record, Figure 5 is presented for clarity.

Record Arrangement

There are three basic areas within each indexed sequential data set: (1) the index area, (2) the prime area, and (3) the overflow area. As was pointed out earlier, the index area provides the facility of "direct" access and the overflow area obviates the need to remake the

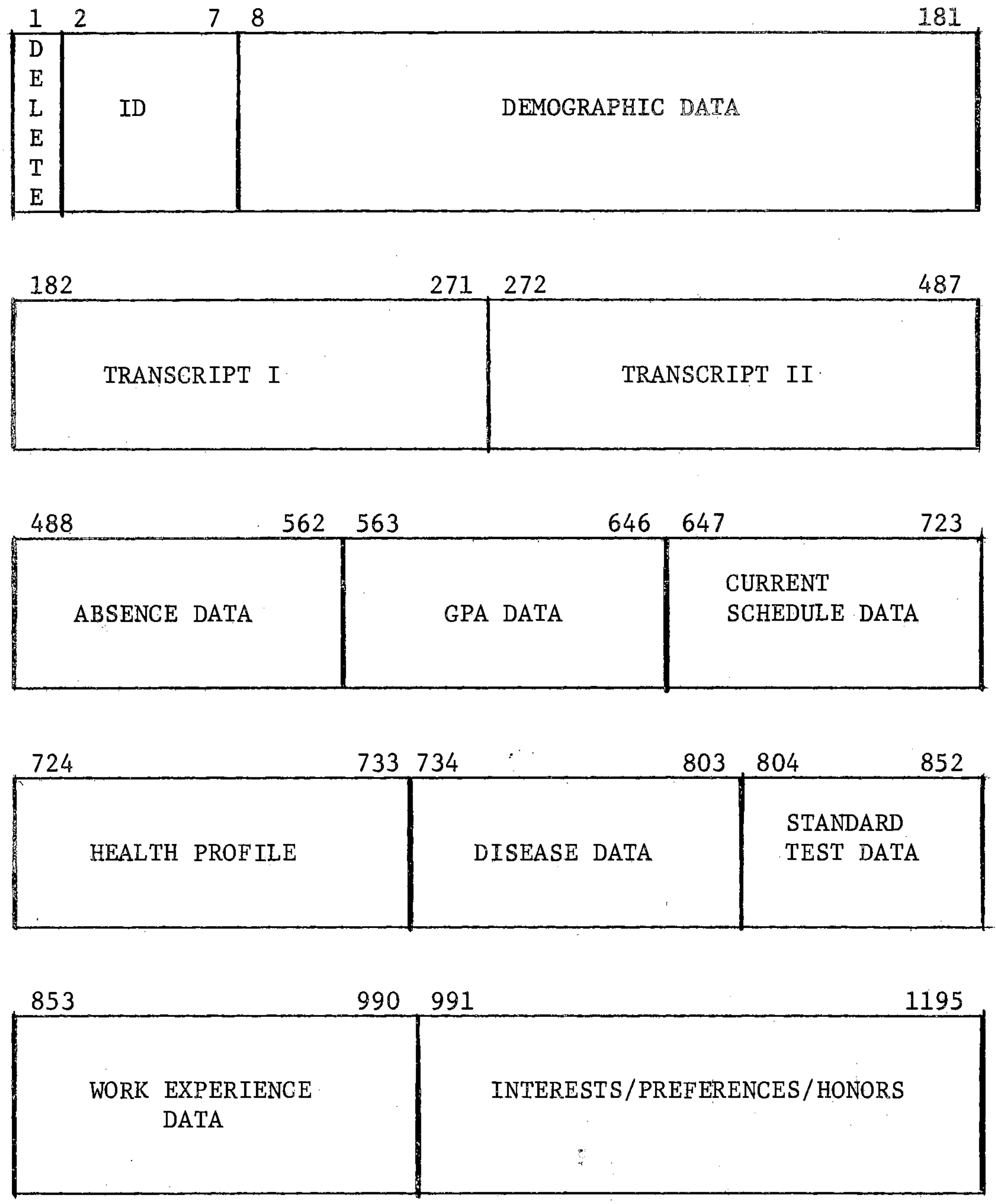


Figure 5. Map of the Student File Record.

TABLE XVI

STUDENT FILE RECORD

Field	Length	Positions
Delete Character	1	1
Student identification number	6	2-7
Student name	23	8-30
Sex	1	31
Race	1	32
Birthdate	6	33-38
Home address		
House number	4	39-42
Street direction	2	43-44
Street name	11	45-55
Home phone	7	56-62
Current grade	2	63-64
High school track	1	65
School code	1	66
Church code (link)	2	67-68
School transfer	15	69-83
Year graduated: 6th	2	84-85
Year graduated: 8th	2	86-87
Father's name	20	88-107
Father's employer	15	108-122
Father's business phone	11	123-133
Mother's name	20	134-153
Mother's employer	15	154-168
Mother's business phone	11	169-179

TABLE XVI (Continued)

Field	Length	Positions
Doctor code (link)	2	180-181
Course 1 (Transcript I)		
Number (link)	4	182-185
Honor points: grade 7	1	186
Honor points: grade 8	1	187
Courses 2-15 (Transcript I)	6 each	188-271
Course 1 (Transcript II)		
Grade	1	272
Number (link)	4	273-276
Honor points: 1st semester	1	277
Honor points: 2nd semester	1	278
Earned credits	2	279-280
Courses 2-24 (Transcript II)	9 each	281-487
Cumulative absences	3	488-490
Weekly absences	2 each	491-562
Grade 7 GPA data	4	563-566
Grade 8 GPA data	4	567-570
Academic Area GPA data	4	571-574
Junior High GPA data	4	575-578
Senior High GPA data	4	579-582
Semi-semester GPA data (16 of them)	4 each	583-646
Course 1 (Current schedule)		
Period of the day	1	647
Course number (link)	4	648-651
Location	3	652-654
Teacher number (link)	3	655-657
Courses 2-7 (Current schedule)	11 each	658-723

TABLE XVI (Continued)

Field	Length	Positions
Health Profile		
Eyes	1	724
Ears	1	725
Teeth	1	726
Growth	1	727
Heart	1	728
Respiration	1	729
Blood	1	730
Skin	1	731
Weight	1	732
Feet	1	733
Disease 1		
Disease code (link)	2	734-735
Date of disease	4	736-739
Date of vaccination	4	740-743
Diseases 2-7	10 each	744-803
Kuhlman-Anderson test	3	804-806
SRA Test		
Parts 1-13	2 each	807-832
Composite score	2	833-834
DAT Test: Parts 1-2	2 each	835-852
JOB 1 (Work Experience)		
Job description	20	853-872
Months of experience	2	873-874
Status	1	875
JOBS 2-6 (Work Experience)	23 each	876-990
IPH 1 (Interests, Preferences, Honors)		
Description	40	991-1030
Type code	1	1031
IPH 2-5 (Interests, Preferences, Honors)	41 each	1032-1195

TABLE XVII

COURSE FILE RECORD

Field	Length	Positions
Delete character	1	1
Course number (search key)	4	2-5
Course name	12	6-17

TABLE XVIII

TEACHER FILE RECORD

Field	Length	Positions
Delete character	1	1
Teacher ID number	3	2-4
Teacher name	10	5-14

TABLE XIX

DOCTOR FILE RECORD

Field	Length	Positions
Delete character	1	1
Doctor ID number	2	2-3
Doctor's name	13	4-16
Doctor's phone	7	17-23

TABLE XX

DISEASE FILE RECORD

Field	Length	Positions
Delete character	1	1
Disease ID number	2	2-3
Disease name	15	4-18

TABLE XXI

CHURCH FILE RECORD

Field	Length	Positions
Delete character	1	1
Religion ID number	2	2-3
Religion name	15	4-18

TABLE XXII

PRIVILEGED DATA FILE RECORD

Field	Length	Positions
Delete character	1	1
Counselor identification	2	2-3
Student identification	6	4-9
Overflow digit	1	10
Line 1, 2, or 3	80	11-90

file when additions and/or deletions are imposed on it. The prime area is where the records are placed when the data set is created. During creation the prime records must be placed in ascending sequential order based on its search key. When records are inserted after creation, the logical ascending key sequence is maintained but some of the records end up in the overflow area.

The record arrangement, then is a logical sequence determined by the ID number of the record for the particular data set.

File Arrangement

All files are to be placed on the disk type of direct access storage device. Data on the IBM 2314 disk pack are organized into cylinders, tracks, and bytes. Each track can contain a maximum of 7,294 bytes, each cylinder is a vertical array of twenty tracks, and each disk pack contains a concentric array of two-hundred cylinders. For each ISAM file, there are three levels of indices; track, cylinder, and master. The first two are always present and the last is optional and is used when the cylinder index becomes too large for an efficient search. The track index is located on the first track of each cylinder of the prime area and its contents point to the data records in the cylinder. The cylinder index entries point to tracks containing track indices and may be placed in an independent area in the disk system or may be placed on a different type of direct access storage device.

The prime area must consist of contiguous tracks within a cylinder and may or may not (depending on the operating system) be placed on several non-contiguous cylinders and can span one or more disk packs.

The overflow area may be part of each cylinder and/or it may be placed in an "independent overflow area." The advantage of the "cylinder overflow area" is that fewer seeks are required to locate the overflow data while the disadvantage is unused space (28:32-34).

The physical placement of these areas within the data set and the placement of the data sets within the disk system* are defined by the user when the files are created. The placements depend on the physical location of available space and the policy standards of the computer installation.

Summary

This chapter answers the first six of the seven questions posed in Chapter I. Some sixty-seven variables were identified to adequately satisfy the needs of a student fact retrieval system for the public school counselors in Bartlesville. In addition to the student data, provisions were made to let the counselor store any additional information about the student in a privileged data file. The file is considered privileged since only an individual counselor has access to it. The data items of the student record were designed to implement the "branching file structure" discussed in Chapter III. This resulted in a considerable saving of disk storage space since some of the recurring data values were placed in five "table data sets." These were course, teacher, doctor, church, and disease names. The records of each file were arranged to satisfy the requirements of the indexed sequential access method. The data base consists of seven data files; student, course, teacher, doctor, disease, church, and privileged data. Each

---

*The IBM 2314 disk system contains eight disk drives, each containing one disk pack.

file in the data base is arranged contiguously by track with the cylinders of data placed according to the availability of space (IBM's OS operating system is assumed). Two types of on-line mass storage media were available for retrieval purposes, viz., magnetic tape and magnetic disk. The disk was chosen because of the need for direct (random) access of the student fact data in a real-time environment.

The answer to the seventh question is discussed subsequently in Chapter V.

CHAPTER V

FILE MAINTENANCE

Introduction

During the course of using the data retrieval system of TGISS, there will certainly be a need to make alterations to the data base. These alterations are called file maintenance. The file maintenance procedures needed for the proposed data base fall into two categories based on changes which are either external or internal.

External Changes

The ongoing activities of the students, teachers, counselors, etc., constitute the external changes. The student may change his name, the school he attends, his address, etc. Time perpetually changes the student's age. The student may make changes or additions to his interests and preferences. And so on. Moreover, all variables of the data base are susceptible to change due to errors which may have been introduced. However, the greatest volume of changes are due to keeping the contents up to date.

Anticipated changes to the data base are:

1. Student additions and deletions to the student data set.

2. Additions and deletions to the course, teacher, doctor, disease, and church files.

3. Alterations to the demographic section of the student file.
4. Additions to the transcript.
5. Nine week GPA update.
6. Insertion of the new course schedule
7. Alterations of the health profile.
8. Additions to work experience.
9. The addition of standardized test scores.
10. Additions to interests, preferences and honors.
11. Additions to the health record.

A File Maintenance System

For the purposes of design it was convenient to divide the input to the file maintenance procedures into two categories: (1) predictable and (2) unpredictable. The predictable data are those which occur at known points in time. Some of these data are created at random but are batched and entered into the system at scheduled times. The points in time when the data are created are the beginning of the year, the end of the year, the beginning of the semester, the end of the semester, the end of a nine week period, and the end of a week. The data that fit into these categories is discussed subsequently.

1. Beginning of a school year. The data in this group includes Health Profile changes, Work Experience additions, and Standard Test additions.

2. End of a school year. This group includes Transcript I entries, and grade 7 and 8 GPA.

3. Beginning of a semester. The current schedule.

4. End of a semester. This group consists of the Transcript II entries, and the Academic Area, Junior High, and Senior High GPA updates.

5. End of a nine-week period. Semi-semester GPA.

6. End of a week. Absences.

The so-called unpredictable data are also created at random but they are batched and processed against the data base at convenient times during the annual cycle. These data include the demographic, disease, and IPH data of the student data set, and all table data sets.

Figures 6-8 depict the logic of the file maintenance system which is broken down into three runs. Run 1 (Figure 6) illustrates how the change data are converted from the source document to a form which is readable by the computer system hardware. This is a two step process where the detail cards are (1) keypunched and (2) verified. The verification step is the first "closed loop" in the several error control feedback loops. If an error is made in step 1 of this run, the verifier is likely to detect it and therefore "feeds" this information back to the keypunch step for correction.

After the data are punched they must travel the long distance between Bartlesville and Stillwater, Oklahoma, and must pass through several hands. For this reason, the data will be grouped into small batches and a control total and card count will be created for each of the batches. The small batch was recommended in order to facilitate locating the error when one is detected. The control total will be a summation of the particular ID number. The purpose of these totals is to determine whether or not any of the data have been lost and serves as another feedback loop for error control. See Run 2 (Figure 7).

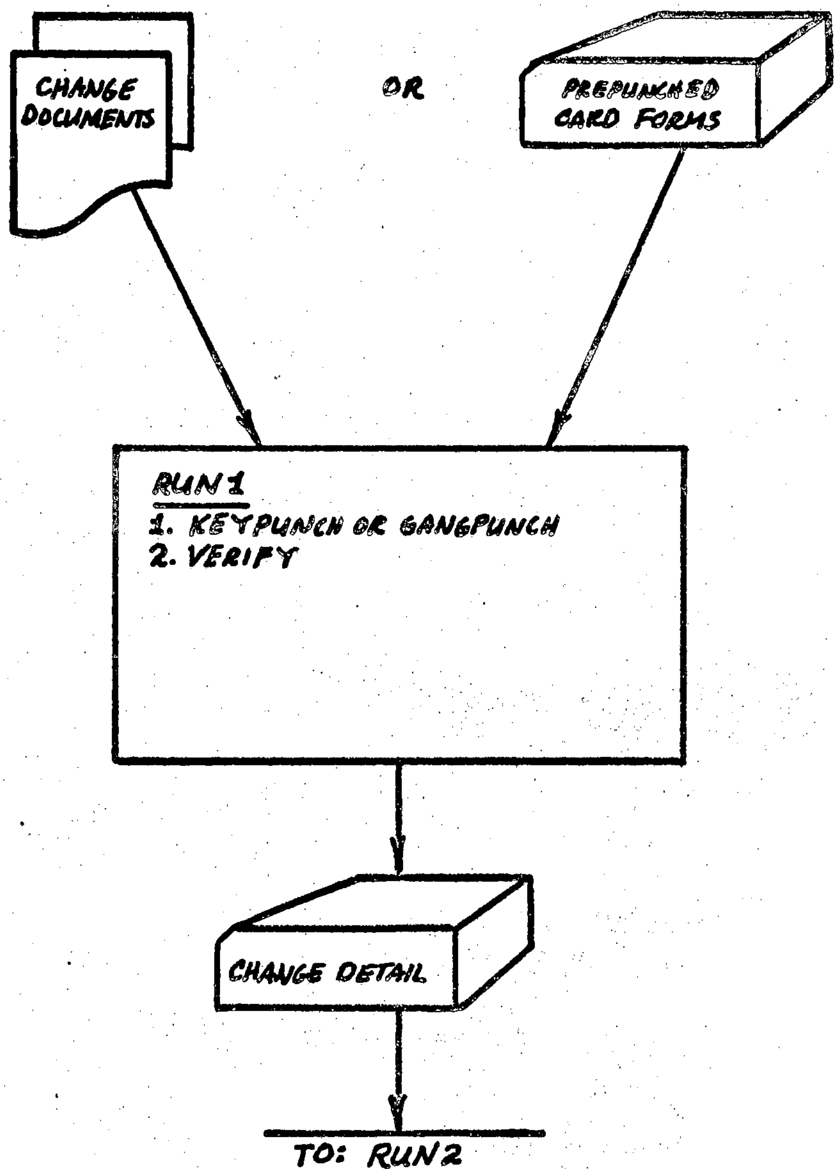


Figure 6. File Maintenance, Run 1

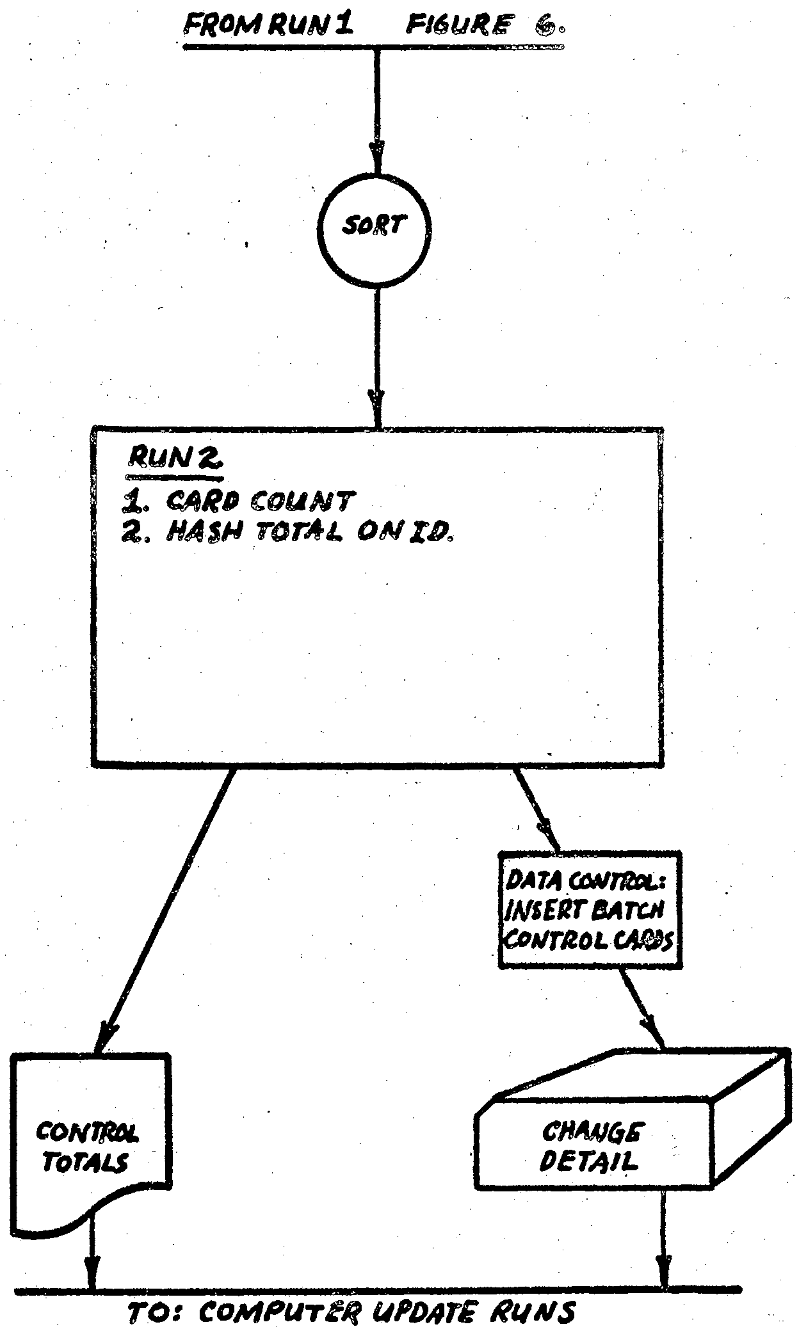


Figure 7. File Maintenance, Run 2

Run 3 (Figure 8) represents the file maintenance operations. It is a generalized computer program to be written to perform the following functions:

1. Add or delete a student record.
2. Add or delete a Table Data Set record.
3. Change "unpredictables" in the student record.
4. Change "predictables" in the student record.
5. Change a Table Data Set record.
6. Convert the letter grade to honor points.
7. Compute grade point averages and cumulative weekly absences.
8. Count the cards in the batch.
9. Sum the ID number.
10. Check for an all numeric ID.
11. Check for the presence of a vadid card code.
12. Prepare an error listing.

The file maintenance logic is shown in Figure 9. The logic was designed with PL/1 in mind. For example, the READ with SET option allows a card to be read into a buffer and the format of its contents to be "set" based on the card code in columns 78-80 (Figure 10). This option eliminates the need for assigning a separate formatted storage area for each card code. Another feature implied by the logic is the fact that PL/1 ignores the closing of an already closed file and the opening of an already open file. Please note that the logic diagram is not detailed and that it is intended only to show a general logic flow. Tables XXIII and XXIV serve to clarify the error types and key words, respectively, of Figure 9.

The card data formats of the input to the file maintenance system

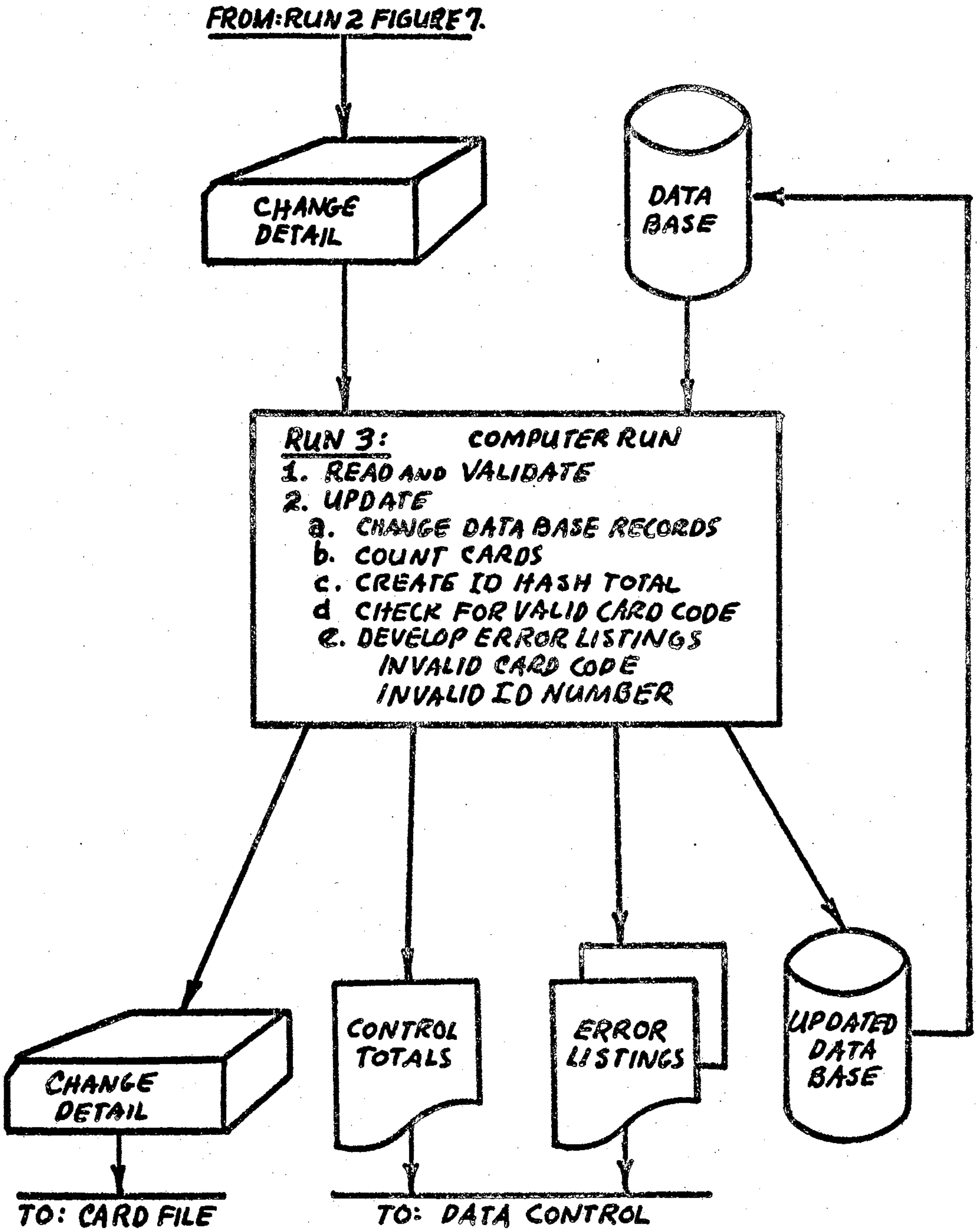


Figure 8. File Maintenance, Run 3

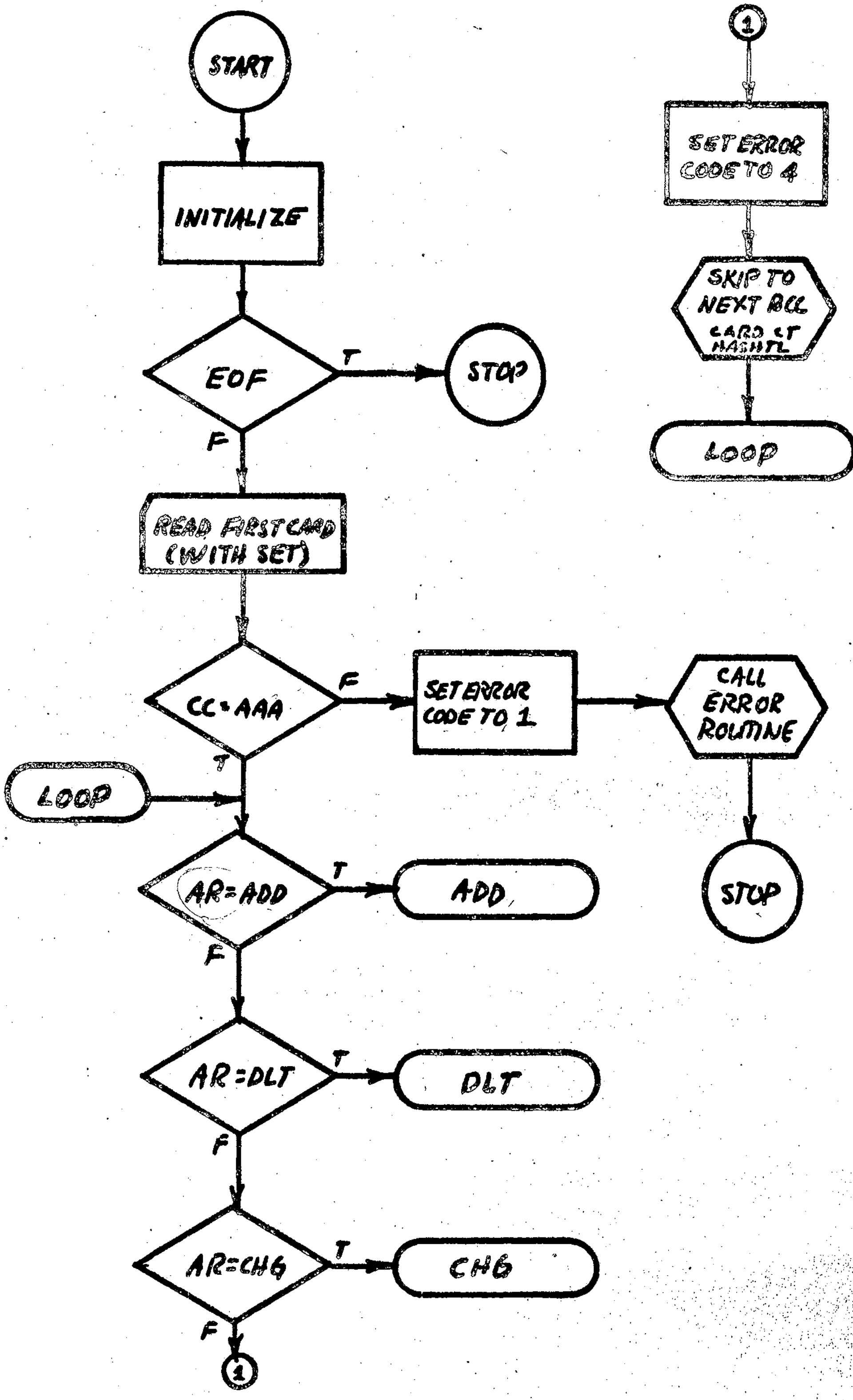


Figure 9. Logic of a File Maintenance System

Figure 9. (Continued)

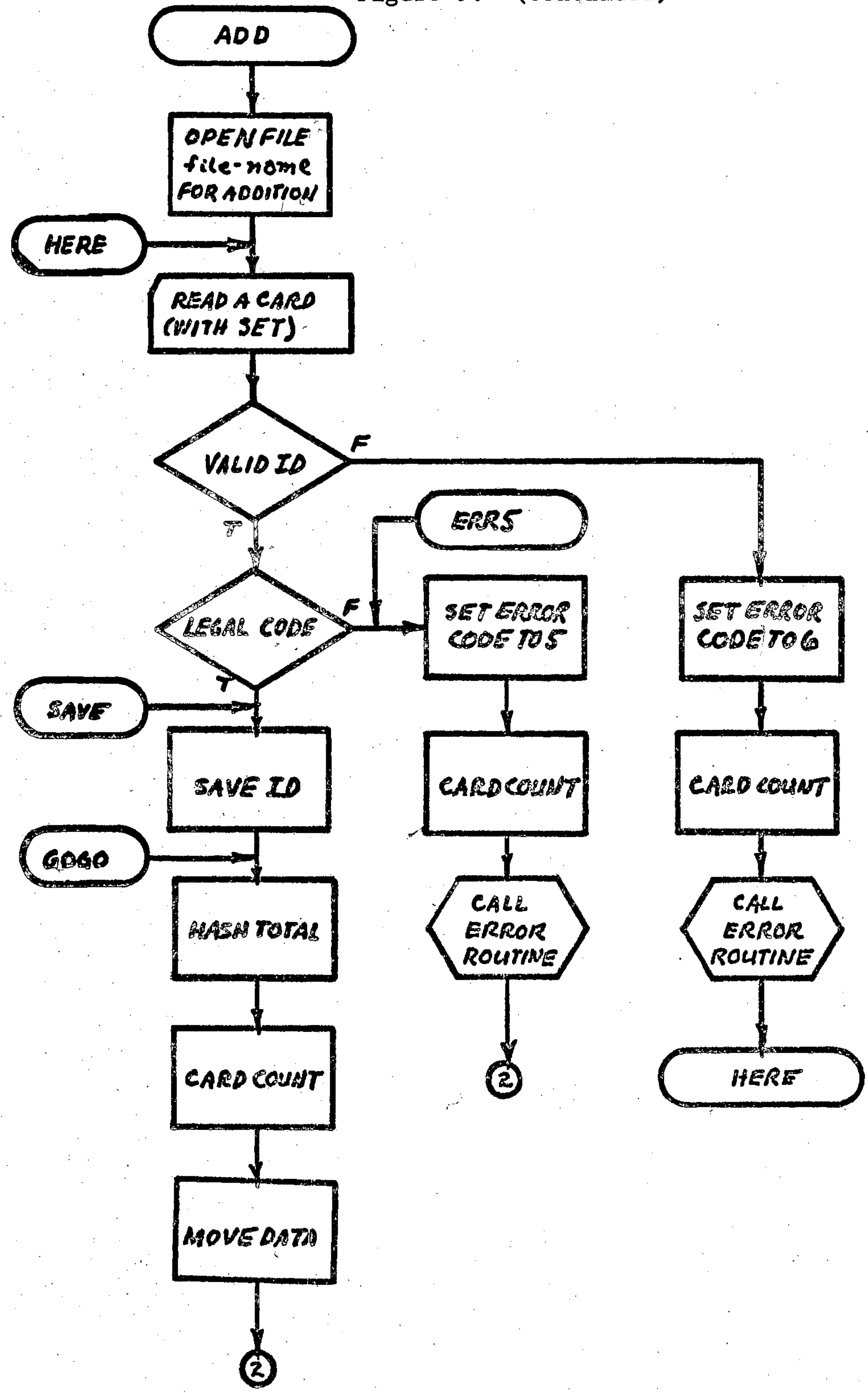


Figure 9. (Continued)

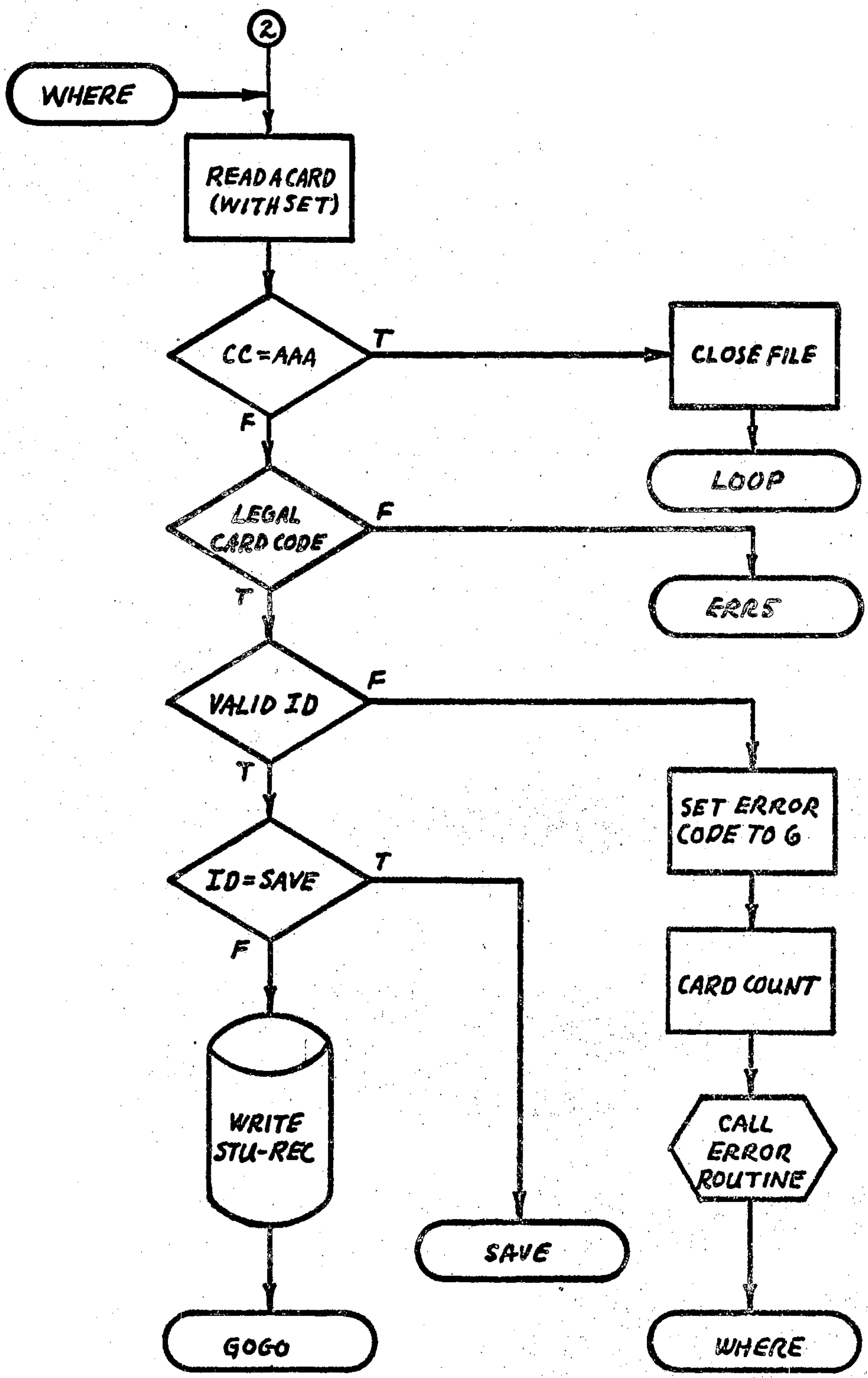


Figure 9. (Continued)

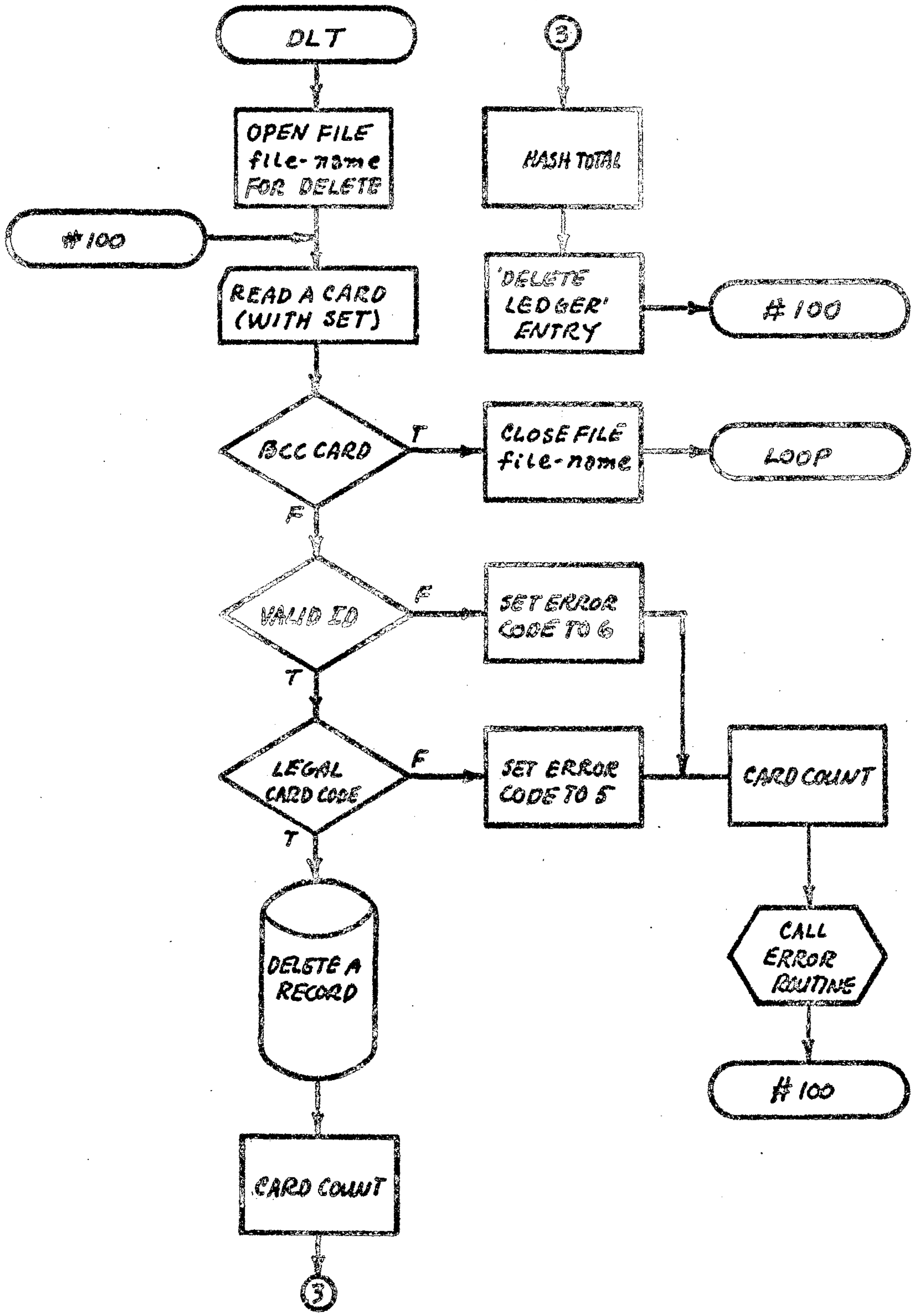


Figure 9. (Continued)

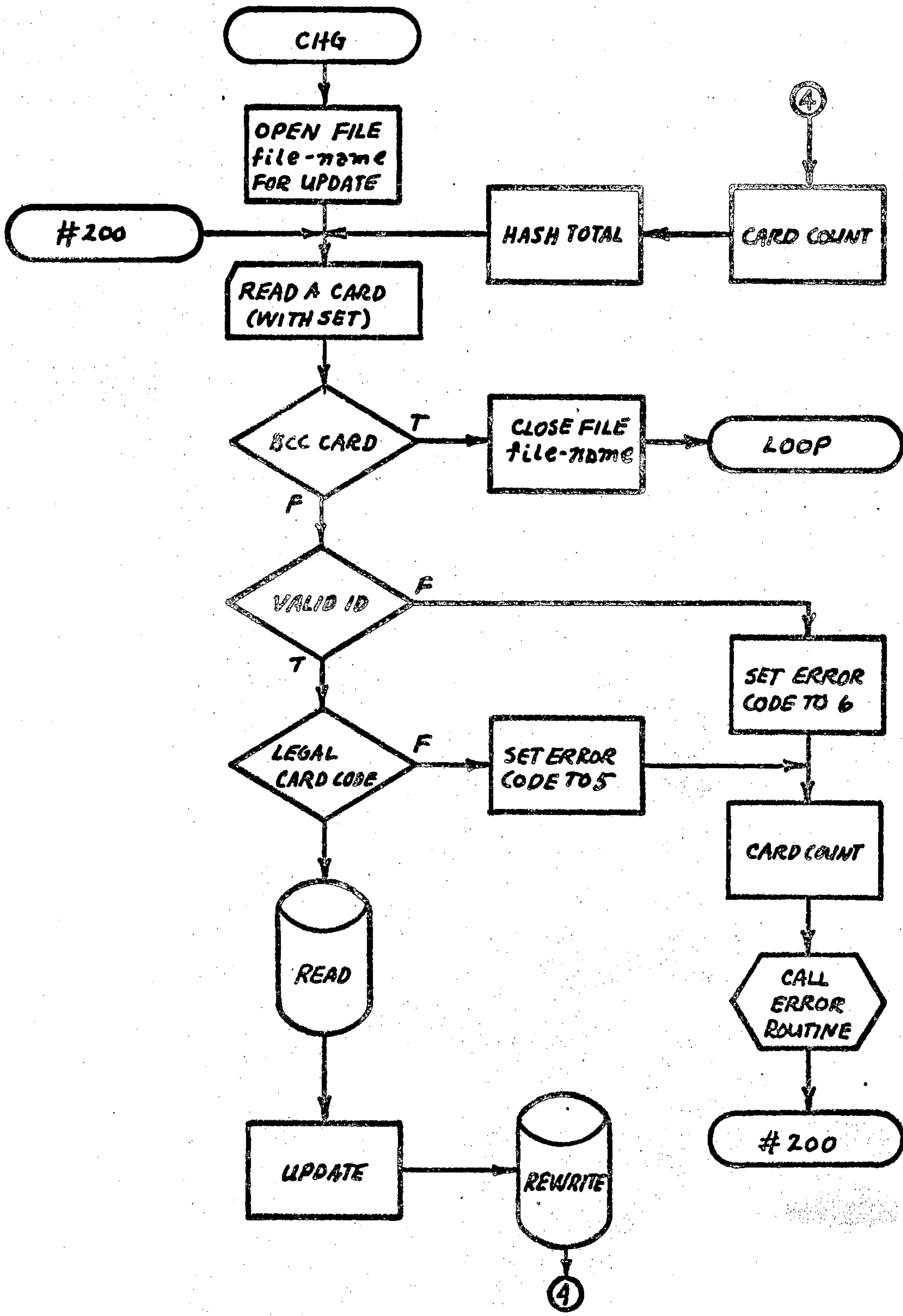


TABLE XXIII

ERROR CODE AND LISTING DATA

Error Type	Meaning
Error 1	First card must be a Batch Control Card.
Error 2	Not Used
Error 3	Not Used
Error 4	Not Used
Error 5	Illegal card code
Error 6	Invalid ID in batch XXX, dated xx xx xx.

TABLE XXIV

KEYWORD LIST FOR FIGURE 9

Acronym	Meaning
ADD	Add a Record
AR	Action Request
BCC	Batch Control Card
CC	Card Code
CHG	Change
DLT	Delete a Record
EOF	End of File
ID	Identification Number
STU-REC	Student Record

are shown in Figure 10. Each format is coded in columns 78-80 with codes 010 through 260 inclusive. There is a ten digit interval between codes. Note that some codes contain an "n" in Column 80. This means that Column 80 may contain 0, 1, 2, 3, ... N-1, where N is the number of cards in the group. For example, if a student had had three diseases (N=3; n=2), the card codes would be 040, 041, and 042. The purpose of this is to maintain the same order throughout as was recorded on the source document. In some cases it was possible to assign all repeating groups to the same card, for example, card code 150 (Figure 10).

The card formats of card codes 010 through 130 are also the formats used for the input to the student data set creation. Card code 140 is used to record the "unpredictables" and the "predictable" work experience. Card codes 150 through 210 are used to record the other "predictables." The Table Data Set updates utilize card codes 220 through 260.

Card format AAA is the Batch Control Card (BCC) which is placed in front of each batch (Figure 11). The purpose of the BCC is to instruct the file maintenance program what to do with the detail data. The instructions are given in the form of an Action Request (columns 10-12). The BCC also indicates, to the file maintenance program, which file is affected (columns 15-17). Table XXV exhibits the list of Action Request codes and data set codes. The BCC also contains the date on which the batch was created and the batch number, both to be used for documentation control purposes. Columns 3-8 of the BCC contains 'UPDATE' which is used to help the operators identify the card.

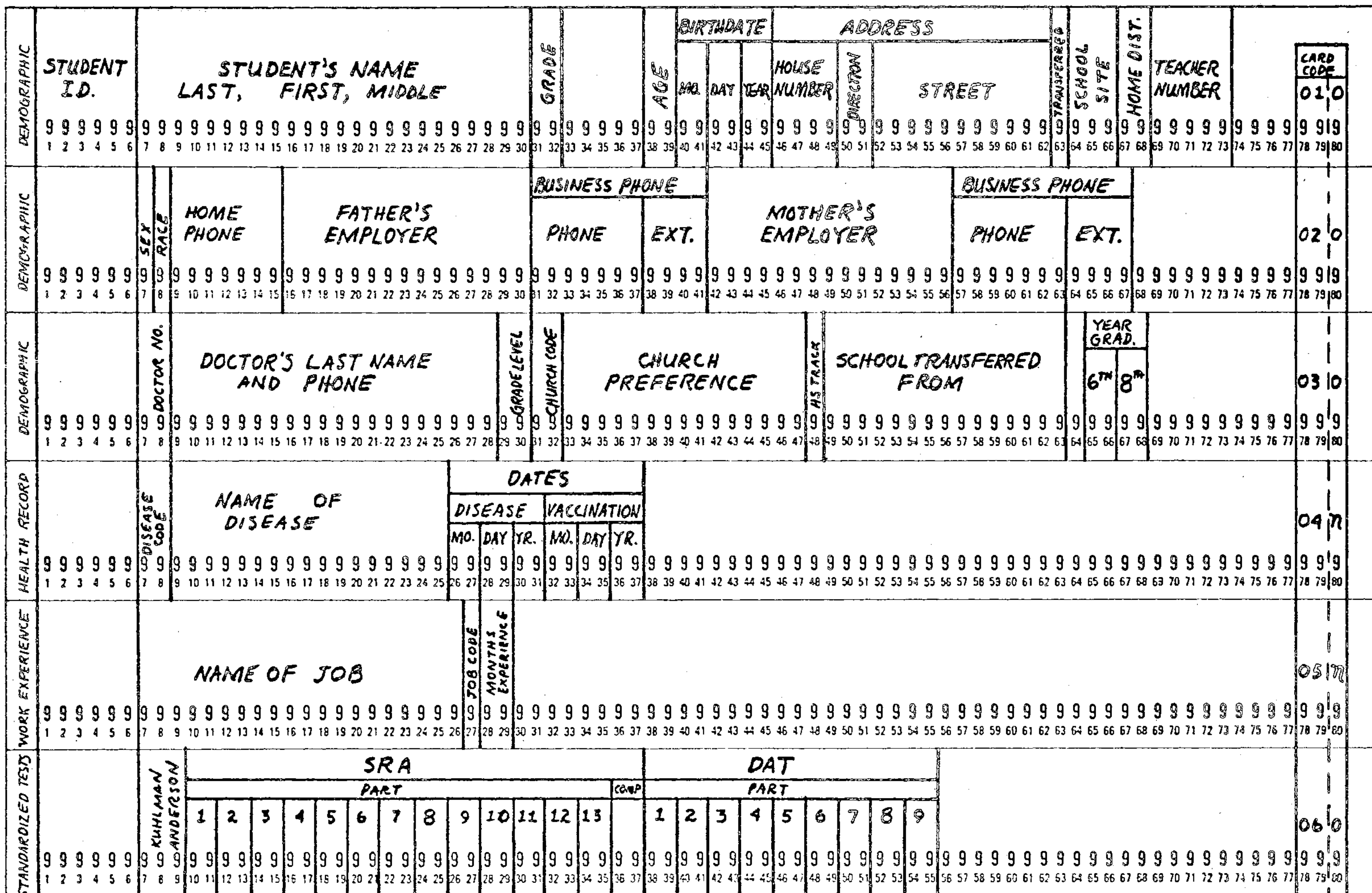


Figure 10. File Maintenance Input Data Formats

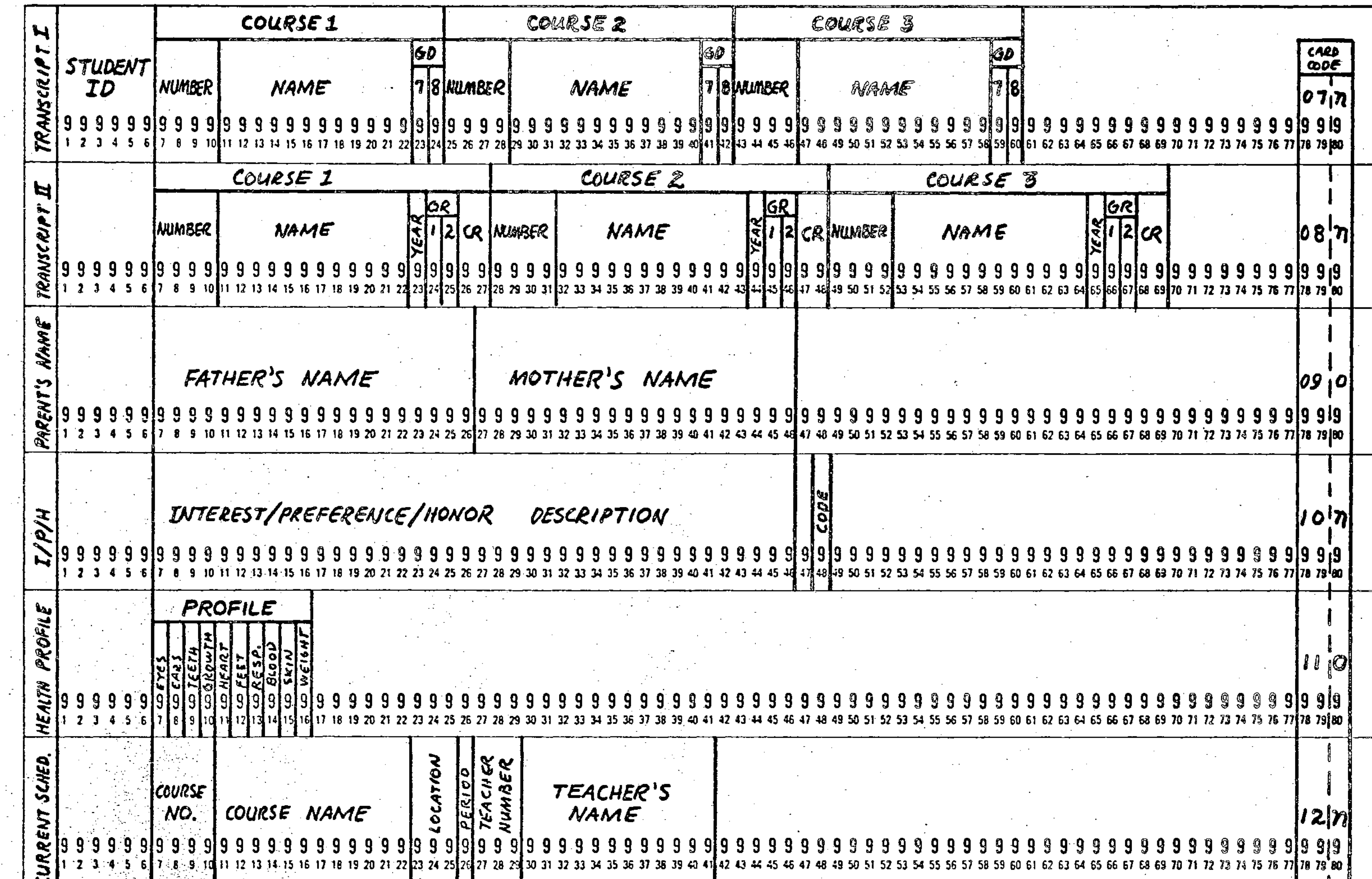


Figure 10. (Continued)

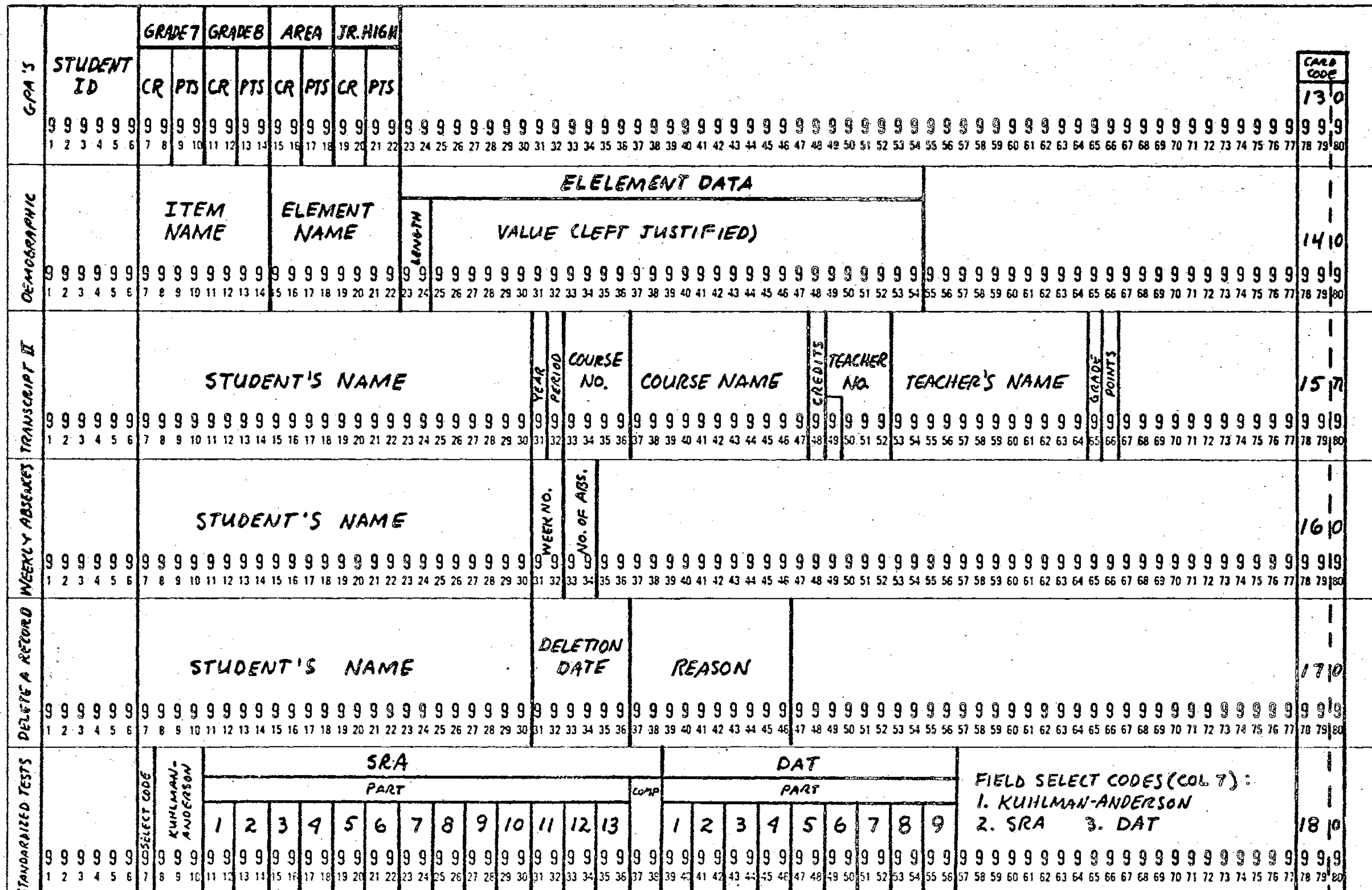


Figure 10. (Continued)

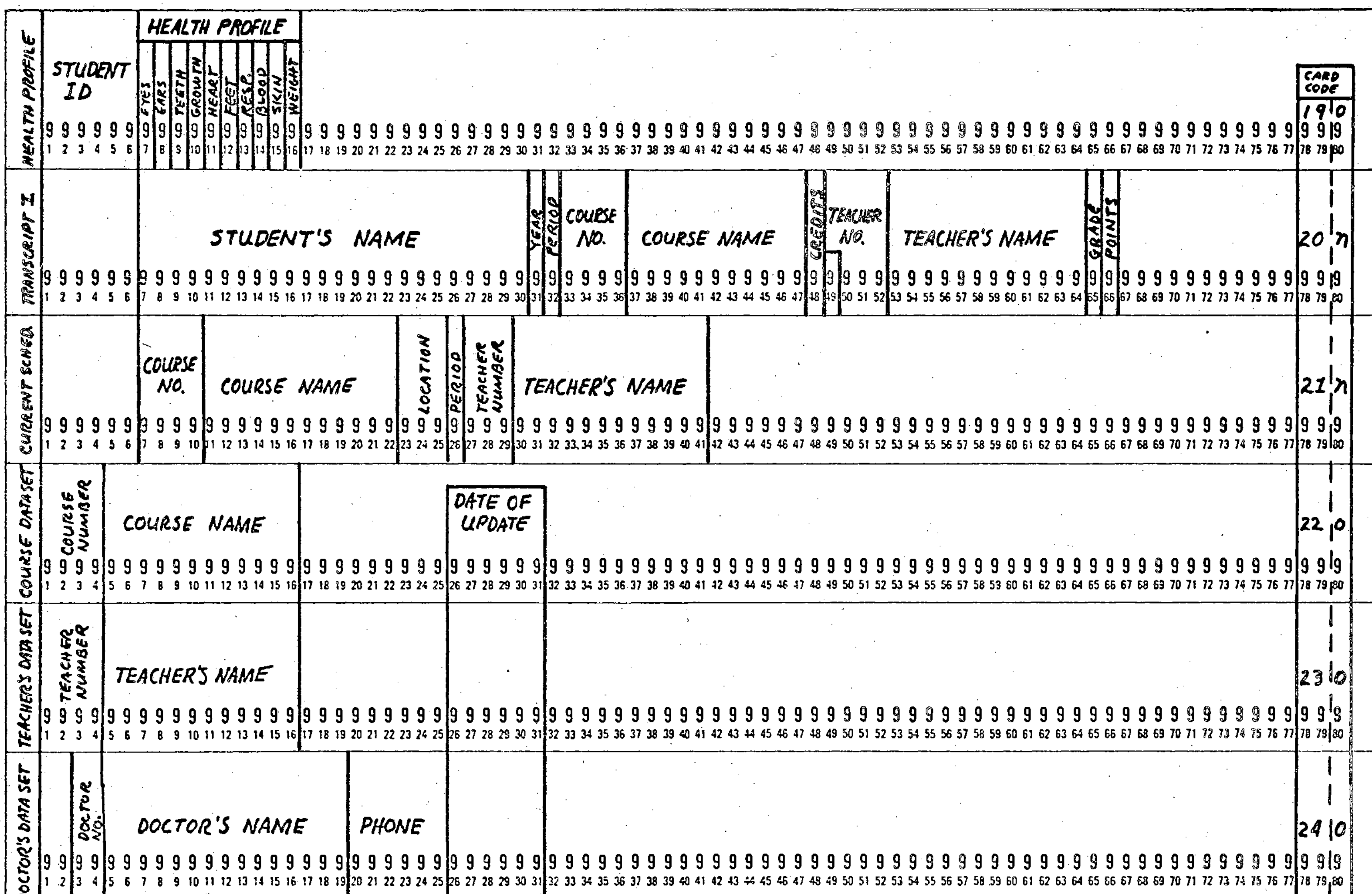


Figure 10. (Continued)

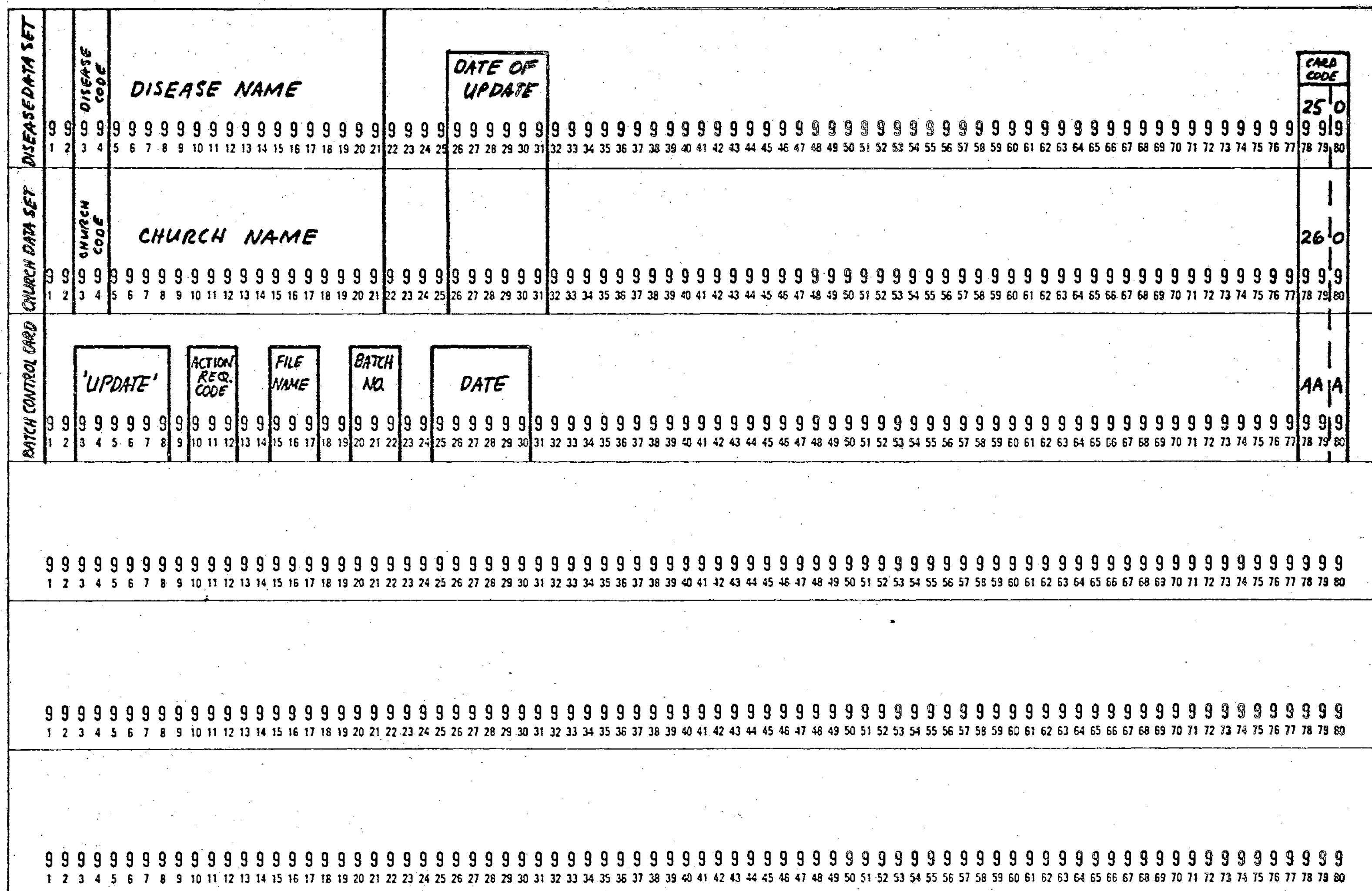


Figure 10. (Continued)

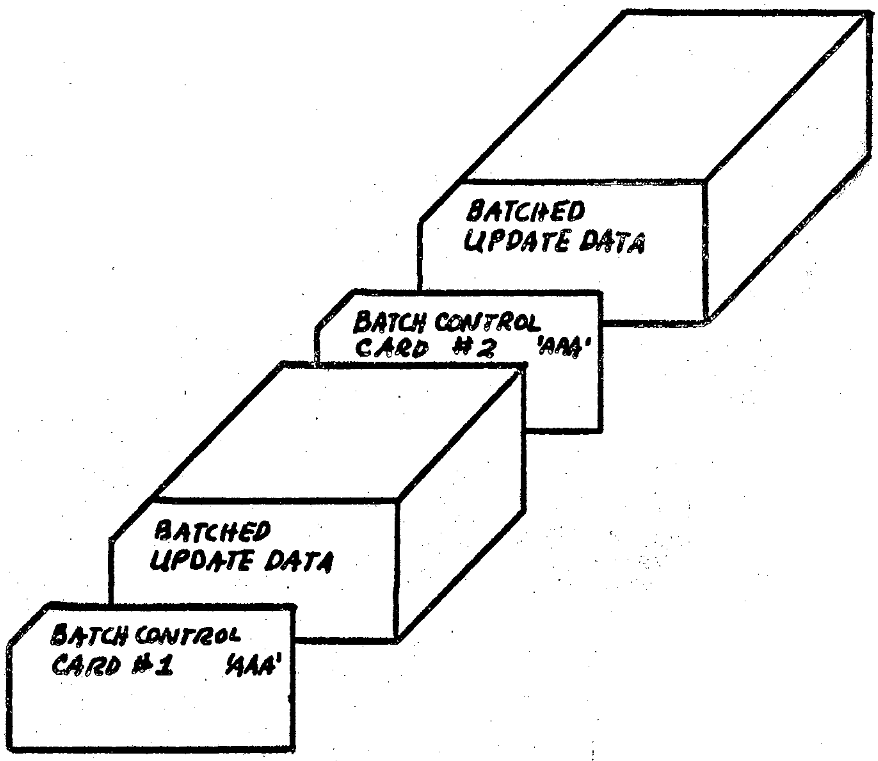


Figure 11. Batch Set-Up

TABLE XXV

ACTION REQUEST AND FILE CODES

Action Request Code	Function	File Code	File Name
ADD	Add a record to the data set	STU	Student File
		CRS	Course File
DLT	Delete a record from the data set	TCH	Teacher File
CHG	Change an element, group, etc., in a record	DOC	Doctor File
		DIS	Disease File
		CHC	Church File

Data Collection

The collection of the data for the file maintenance function is performed by completing the applicable form. There are two kinds of data collection devices: (1) a preprinted form which is completed by the counselor, teacher, or student and (2) the pre-punched card form. Figures 12.1 through 12.9 (see Appendix) illustrate the preprinted forms which are used to initiate data for the addition of a student record. These forms are the same as those used for the data base creation. Figures 13.1 through 13.7 (see Appendix) show the preprinted forms and card forms used to enter the remaining update data.

An example of a pre-punched card form is shown in Figure 13.2. Columns 1-64 and 78-79 have the indicated items pre-punched and will appear in "natural language" across the top of the card. At the end of the semester the teacher will enter the student's grade and sign the card. The teacher will also be asked to batch and count the cards according to letter grade. That is, all of the A's will be batched and counted, all of the C's, and so on. These batch totals will be sent with the data for control purposes. The batched data will be sent to Data Control where the grades and honor points will be gang-punched into the cards. The same procedure will be used for the other pre-punched card forms.

Internal Changes

The underlying property of the external changes discussed above is that the "transaction" effects an individual data set or record contents. If a student changes his family doctor, only the contents of this element is changed. If a student is added to the file, only the contents of one of the previously allocated record storage positions is changed. But, the format of the individual record, data set, and the data base remain the same.

Internal changes reflect a change in design due to (1) the inclusion of new data elements not previously considered important, (2) the deletion of data elements or data sets no longer important, (3) the formation of new data sets based on an analysis of recurring data elements such as the course names, and (4) a change in philosophy such as switching to a new and better query language or changing the basic computer hardware. All of these changes require some change in format of the data set and/or data base as well as the contents. Internal changes may also necessitate the collection of additional data and will certainly require a regeneration of the data base.

Because of the "mysterious" nature of the internal type of change, no attempt was made to define a model to implement a change. This type of change is made by recycling through the procedures previously discussed.

Privileged Data Set

The privileged data belonging to each counselor are placed into a single data set which is keyed on counselor and student identification, and overflow digit (nine numeric characters). The data are arranged in groups called "pages" and "books." Each counselor has a book with a set of pages equal in number to his counselee count. Each page contains three lines of eighty characters which the counselor enters by using FASTER's update feature. If the counselor should need additional lines, he can record up to nine by utilizing the overflow characters.

This "overflow" system was chosen because it was determined that the counselor would rarely need more than three lines for each student.

Summary

This chapter has presented a model for the maintenance of the data base described in Chapter IV. Two kinds of data alterations were defined and discussed.

One type of alteration is due to "external changes" of the student data variables. A model file maintenance system was presented for the implementation of the external change. The system model includes the procedural logic, a computer run logic, input data formats, and the data collection forms. The need for error control was emphasized. The fundamental characteristic of this type of alteration is that only the contents of the data sets change while their formats remain constant.

The other type of alteration is the "internal change" resulting from the addition and/or deletion of data elements or data sets, or a change in operating philosophy. The difference between this type of change and the external change is that both the format and the data base contents are effected.

CHAPTER VI

SUMMARY

The Problem and Approach

As result of the increased complexity of the twentieth century society, timely and comprehensive information relevant to individual achievement has become very important but also an increasingly difficult commodity to obtain. It has become apparent, for example, that the counselor and teacher should not interpret test scores by themselves, but should correlate them with other data available on the student and his surroundings. Thus, the need for more and better information about the student; and more sophisticated methods of interpretation of the data has inspired the design of information systems which utilize present-day computing capabilities.

One such system is project TGISS, a total guidance information support system. TGISS was conceived to provide support in the areas of information retrieval, vocational awareness, diagnosis for remedial education, and vocational-educational guidance. All of these areas require basic information about the student. The purpose of this study was to design a data base of student fact data for TGISS.

The investigation began by assuming (1) that there is a body of information which is vital to the guidance function, (2) the information is hierarchial such that it may be placed into a structured system, and (3) a digital computer can be used more effectively (in terms of speed and accuracy) than any present-day human system.

The investigation set out to answer several questions related to the data base design. These questions were:

1. What constitutes an adequate set of student fact data for a secondary school system?

2. How should the data records be structured to facilitate retrieval and processing?

3. How should the data file be structured to facilitate retrieval and processing?

4. How many data files should be included in the data base?

5. How should the data files be organized within the data base?

6. What is the role of the storage medium in the organization of the data base?

7. How should the resulting data base be kept up-to-date (file maintenance)?

The above questions were answered by (1) reviewing the literature related to data base design and information systems, (2) an analysis of the information needs in the Bartlesville (the geographical location of TGISS) secondary schools, (3) determining how student fact data are used in the counseling function, and (4) studying the present information system to determine points and frequency of data origin, and data flows, in order to outline a file maintenance system.

The Findings

Two work conferences between administrators, counselors, and researchers were used to identify the variables which were included in the data base. The purpose of this step was to answer Question #1, pertaining to an adequate set of data. The result was the identification of 67 variables in eleven categories. These categories were (1) General (name, age, sex, etc.), (2) Transcript of courses completed, (3) Absences, (4) Grade point averages, (5) Course schedule, (6) Health profile (current health status), (7) Health record (historical health status), (8) Standardized test scores, (9) Work experience, (10) Interests, preferences, or honors, and (11) Privileged data as entered by the counselor. A detailed listing is given in Chapter IV.

Question #2, concerning record structure (Data Item Design section of Chapter IV), was answered in terms of (1) the retrieval language and (2) storage space. The retrieval language was FASTER (Filing and Source data entry Techniques for Easier Retrieval, an IBM Type 3 language) which requires fixed length records. The arrangement of the data items within the record was arbitrary except for the first character which had to be a "delete" character as required by FASTER, and the search key (student ID number) which had to be in six contiguous positions. However, the data items were grouped by functions. For example, all grade point data were grouped together and given the name Grade Trends. These groupings were made to make it simpler, logically, to locate the data for retrieval. The resulting data item contents are shown in Tables I through XII.

The storage space aspect involved a space reduction. This reduction was accomplished by abbreviating and coding certain data item elements. Coding, like abbreviation, serves to replace a larger data element by one which is much shorter. Coding especially paid off where there was an excessive amount of redundant data values for those elements which were not unique to the individual. An example is course names, where the same name appeared thousands of times throughout the file in transcripts and course schedules. The code served as a link to another data set where the course name appeared only once. With this scheme, the student record was reduced from 1,938 characters to 1,195 characters; a savings of 743 characters per student.

Question #3 was answered by the retrieval language, FASTER, which requires the indexed sequential data set organization. With this method, the data records were recorded sequentially in the prime data area. Subsequent retrieval can be sequential or random. If random, the data record is found by a table look-up in the index file created when the data set was created. Subsequent additions to the file may cause a departure from the original sequential organization since the record to be inserted may cause a record to be placed into an overflow area. The logical sequence is maintained, however, by linking the overflow records to each other and to those in the prime area.

The proposed data base consisted of seven data sets (Question #4). The names of these files are:

1. Student File,
2. Course File,
3. Teacher File,
4. Doctor File,
5. Disease File,
6. Church File,
7. Privileged Data File.

Files 2 through 6 resulted from the storage space reductions discussed above. These files require an estimated 8,500 characters of storage. However, recall that the addition of these files resulted in a decrease of 743 characters from each student record. For a 4,215 student data base, this amounts to a gross reduction of 3,131,745 character positions. Thus the addition of some 8,500 characters was a small price to pay for ever three million characters of file space.

Questions #5 and #6 are highly interrelated and must be discussed together. There were two types of storage media available to project TGISS; magnetic tape and magnetic disk. Magnetic tape allows sequential organization only with sequential processing easily accomplished but with random processing highly impractical. Magnetic disk can implement sequential and random organization with sequential processing as easy as magnetic tape and random processing a highly desirable characteristic. The nature of project TGISS required that random processing be available, hence, the magnetic disk was utilized. Random processing does not imply random organization as evidenced by the index sequential method.

File arrangement on magnetic tape must be strictly sequential and the location of a file on tape would probably be located according to the frequency of processing. Not so for the disk. The file, as an entity, can be placed in any set of contiguous tracks. Consequently, the arrangement of the data sets depended on the physical location of available space and the policy standards of the computer center.

The file maintenance procedures (Question #7) included making two kinds of updates; (1) those due to changes created by the student, teacher, doctors, etc., and (2) those created by improvements and innovations added to the retrieval system. Only the first type were considered in this study.

All elements of the data base are susceptible to change. Chapter

V proposes a model to make these changes. The model includes the procedural logic, a computer run logic, input data formats, and the data collection forms. The need for error control was emphasized. It was pointed out that routinely only the contents of the data sets change while their formats remain constant. Changes to the data set formats are outside the scope of the model.

SELECTED BIBLIOGRAPHY

(1) Abrams, Peter, and Walter Corvine. Basic Data Processing. New York: Holt, Rinehart and Winston, Inc., 1966.

(2) "A Computerized Career-Planning File Leads Students to 'Self-Help'," Journal of College Placement, 28 (December 1967-January 1968), pp. 39, 40.

(3) Anderson, G. Ernest, Jr. "100 Uses for School Processing," Educational Data Processing: New Dimensions and Prospects, ed. Richard A. Kaimann and Robert W. Marker. Boston: Houghton Mifflin Company, 1967, pp. 206-9.

(4) Artandi, Susan. An Introduction to Computers in Information Science, Metuchen, J. J.: The Scarecrow Press, Inc., 1968.

(5) Awad, Elias M. Automatic Data Processing, Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1966.

(6) Awad, Elias M. Business Data Processing. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1965.

(7) Bachman, C. U. and S. B. Williams. "A General Purpose Programming System for Random Access Memories," AFIPS Conference Proceedings, 1964, Fall Joint Computer Conference, cited in The Analysis of Information Systems by Charles T. Meadow, pp. 194-8.

(8) Bicknell, John E. "Order from Information Chaos," Minnesota Journal of Education, 46 (May, 1966), p. 25.

(9) Bohn, John J., Jr. and Donald E. Super, "The Computer in Counseling and Guidance Programs," Educational Technology, 9 (March 1969), 29-31.

(10) Carroll, J. B. and A. B. Ellis. Planning and Utilization of a Regional Data Bank for Educational Research Purposes, Cambridge, Mass.: Harvard University Graduate School of Education, 1965.

(11) Cogswell, John F. "Computers in Student Appraisal and Educational Planning," in Man-Machine Systems in Education, ed. John W. Loughary. New York: Harper and Row, 1966, 157-67.

(11) Cogswell, John F. "Computers in Student Appraisal and Educational Planning," in Man-Machine Systems in Education, ed. John W. Loughary. New York: Harper and Row, 1966, 157-67.

(12) "Computer Managed Education," Educational Technology, 8 (May 30, 1968), pp. 16-7.

(13) "Computerized Counseling," Educational Technology, 9 (March 1969), pp. 76-7.

(14) Cooley, William W. "A Computer-Measurement System for Guidance," Guidance: An Examination, ed. Ralph L. Mosher, et al. New York: Harcourt, Brace and World, Inc., 1965.

(15) Dodd, George G. "Elements of Data Management Systems," Computing Surveys 2 (June, 1969), pp. 117-33.

(16) Filing and Source Data Entry Techniques for Easier Retrieval, an IBM Type III program submitted by R. E. GENTRY. New York: International Business Machines, Corp., 1968.

(17) Flanagan, John C. "Project Plan," Third Symposium for Systems Under Development for Vocational Guidance: A report of a research exchange conference, September 14-15, 1967. International Business Machines Corporation, Advanced Systems Development Division, 2651 Strang Blvd., Yorktown Heights, New York, p. 15.

(18) Goldberg, Albert L. and Robert J. Cymbala. "EDP in Education," Audio-visual Instruction, 12 (April, 1967), pp. 349-52.

(19) Grossman, Alvin. "Computers for Counseling and Guidance," Man-Machine Systems in Education, ed. John W. Loughary. New York: Harper and Row, 1966, pp. 168-85.

(20) Harris, Jo Ann. "The Computerization of Vocational Information," Vocational Guidance Quarterly, 17 (September 1968), pp. 12-20.

(21) Helm, Carl. "Computer Simulation Techniques for Research on Guidance Problems," Personnel and Guidance Journal, 46 (Sept. 1967), pp. 47-52.

(22) Hodge, Carl. "EDUCOM," Audiovisual Instruction, 12 (April 1967), pp. 352-5.

(23) Impellitteri, Joseph T. "A Computerized Occupational Information System," Vocational Guidance Quarterly, 15 (June 1967), pp. 262-4.

(24) Impellitteri, Joseph T. "A Computer-Assisted Career Exploration System," National Business Education Quarterly, 37 (Winter 1968), pp. 47-51.

(25) Impellitteri, Joseph T. "Exploring the Potential of Computer-Assisted Vocational Guidance," School Shop, 27 (April, 1968), pp. 78-80.

(26) Impellitteri, Joseph T. The Development and Evaluation of a Pilot Computer-Assisted Occupational Guidance Program: Final Report, Department of Vocational Education, Pennsylvania State University, 1968.

(27) Information System for Vocational Decisions: Second Annual Report 1967-1968, Graduate School of Education, Harvard University, 1968.

(28) Introduction to IBM System/360 Direct Access Storage Devices and Organization Methods. New York: International Business Machines, Corp., 1966.

(29) Johnson, Donald H. "Automated Developmental Counseling System," Third Symposium for Systems Under Development for Vocational Guidance, A report of a research exchange conference, September 14-15, 1967. International Business Machines Corporation, Advanced Systems Development Division, 2651 Strang Blvd., Yorktown Heights, New York, pp. 11-3.

(30) Knuth Donald E. The Art of Computer Programming, Reading, Mass.: Addison-Wesley, 1968.

(31) Laden, H. N. and T. R. Gildersleeve. Systems Design for Computer Applications, New York: John Wiley & Sons, 1963.

(32) Lancaster, F. Wilfred. Information Retrieval Systems, New York: John Wiley & Sons, 1968.

(33) Loughary, John W. "The Computer-Assisted Counselor," NEA Journal, 56 (February, 1967), pp. 22-4.

(34) Loughary, John W. and Murray Tondow. "Computers as Substitute Counselors," Educational Technology, 9 (March 1969), pp. 33-6.

(35) Marker, Robert W. "Educational Information Centers," Educational Leadership, 24 (January, 1967), pp. 367-77.

(36) Meadow, Charles T. The Analysis of Information Systems. New York: John Wiley & Sons, 1967.

(37) Minker, Jack and Jerome Sable. "File Organization and Data Management," Annual Review of Information Science and Technology, Volume 2, ed. Carlos A. Cuadra, New York: John Wiley and Sons, 1967, pp. 123-60.

(38) Minor, Frank., Roger A. Myers, and Donald E. Super. "An Experimental Computer-Based Educational and Career Exploration System," Personnel and Guidance Journal, 47 (February

1967), pp. 564-9.

(39) Poland, C. B. "Advanced Concepts of Utilization of Mass Storage," Information Processing 1965; Proceedings of IFIP Congress 65, New York, 24-29 May. Volume 1, ed. Wayn A. Kalenich. Washington, D. C.: Spartan Books, 1965, pp. 249-54.

(40) Roberts, Tommy L., Charles Kieser, and Ed Forsberg. Theoretical Framework. Technical Memorandum No. 5, The Bartlesville System, Oklahoma State University Research Foundation, Stillwater, Oklahoma.

(41) Salton, Gerald. Automatic Information Organization and Retrieval, New York: McGraw-Hill Book Company, 1968.

(42) Sharpes, Donald K. "Computers in Education," Clearing House, 43 (November 1968), pp. 135-8.

(43) Sippl, Charles J. Computer Dictionary and Handbook, New York: Howard Sams & Co., 1966.

(44) Smith, Thomas W. "Development of a Research-Based and Computer-Assisted Guidance System," Educational Leadership Research Supplement, 25 (May, 1968), pp. 754-60.

(45) Sterling, Theodore D., and Seymour V. Pollack. Introduction to Statistical Data Processing, Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1968.

(46) Tondow, Murray and Mary L. Betts. "Computer-Based Course Selection and Counseling," Journal of Educational Data Processing, 9 (March) 1969, pp. 216-41.

(47) Whitlock, James W. Automatic Data Processing in Education. New York: The MacMillan Company, 1964.

(48) Youst, David B. "The Rochester Career Guidance Project," Educational Technology, 9 (March 1969), pp. 39-41.

APPENDIX

GENERAL INFORMATION

ID NUMBER _ _ _ _ _ _ NAME: ______________________________
Last First Middle

SEX: M Male
F Female

RACE: C Caucasian
L Latin
N Negro
I Indian
O Oriental

ADDRESS: ______________

PHONE: _ _ _ _ _ _ _ _ BIRTHDATE: ______________
Month Day Year

FATHER'S EMPLOYER: _ _ _ _ _ _ _ _ _ _ _ _ _ _ BUSINESS PHONE: _ _ _ _ _ _

FATHER'S NAME: _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ EXTENSION: _ _ _ _ _ _

MOTHER'S EMPLOYER: _ _ _ _ _ _ _ _ _ _ _ _ _ _ BUSINESS PHONE: _ _ _ _ _ _

MOTHER'S NAME: _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ EXTENSION: _ _ _ _ _ _

FAMILY DOCTOR: LAST NAME _ _ _ _ _ _ _ _ _ _ CODE _ _ _ PHONE: _ _ _ _ _ _

CURRENT GRADE LEVEL: _ _ _ CHURCH PREFERENCE _ _ _ _ _ _ _ _ _ _ _ _ _ _

HIGH SCHOOL TRACK: C COLLEGE PREP
G GENERAL
V VOCATIONAL
S SPECIAL

NAME OF SCHOOL ENROLLED IN _ _ _ _ _ _ _ _ _ _ _ _

SCHOOL TRANSFERRED FROM _ _ _ _ _ _ _ _ _ _ _ _ AREA NAME ____________ ____
Code

YEAR GRADUATED FROM 6TH: 19___ YEAR GRADUATED FROM 8TH: 19____

HEALTH RECORD

Code	Name of Disease	Date of Disease	Vaccination Date

Figure 12.1. General Information (Demographic) Form

STUDENT JOB EXPERIENCE

ID NUMBER:_ _ _ _ _ _

NAME:______________________ ______________________ ______________________
Last First Middle

Please list the jobs you have had. Indicate if it was a non-paying, voluntary (V) job or a paying (N) job.

		Circle One		Months Experience
1.	______________________	V	N	______________
2.	______________________	V	N	______________
3.	______________________	V	N	______________
4.	______________________	V	N	______________
5.	______________________	V	N	______________

Figure 12.2 Work Experience Form

ID NUMBER_ _ _ _ _ _ NAME: ____________________ ____________________ ____________________

Last First Middle

STANDARDIZED TEST SCORES

KUHLMAN_ _ _ _

SRA

______	______	______	______	______	______	______	______	______	______
Part 1	Part 2	Part 3	Part 4	Part 5	Part 6	Part 7	Part 8	Part 9	Part 10
______	______	______							
Part 11	Part 12	Part 13							

DAT

______	______	______	______	______	______	______	______	______
Part 1	Part 2	Part 3	Part 4	Part 5	Part 6	Part 7	Part 8	Part 9

INTERESTS/PREFERENCES/HONORS

1. ____ I P H

2. ____ I P H

3. ____ I P H

4. ____ I P H

Figure 12.3. Standardized Test Scores and Interest/Preferences/Honors Form

TRANSCRIPT I - GRADES 7 & 8

ID NUMBER_ _ _ _ _ _ NAME: ____________________
(do not punch)

Course Number	Course Name	Letter 7th	Grade 8th	Course Number	Course Name	Letter 7th	Grade 8th
0101	English 7	_____	_____	0501	Art 7	_____	_____
0102	English 8	_____	_____	0502	Art 8	_____	_____
0201	Social Studies	_____	_____	0551	Instrumental 7	_____	_____
0202	Social Studies	_____	_____	0552	Instrumental 8	_____	_____
0301	Science 7	_____	_____	0561	Vocal Music 7	_____	_____
0302	Science 8	_____	_____	0562	Vocal Music 8	_____	_____
0401	Math 7	_____	_____	0701	Spanish 7	_____	_____
0402	Math 8	_____	_____	0702	Spanish 8	_____	_____
1101	P. E. 7	_____	_____	0801	Home Ec. 7	_____	_____
1102	P. E. 8	_____	_____	0802	Home Ec. 8	_____	_____
				0901	Ind. Art 7	_____	_____
				0902	Ind. Art 8	_____	_____

Figure 12.4. Grades 7 and 8 Transcript Form

TRANSCRIPT - NINTH GRADE

ID NUMBER_ _ _ _ _ _

Course Number	Course Name	Letter Grade First Semester	Letter Grade Second Semester	Credits
0103	English I	____	____	____
0145	Newspaper Staff	____	____	____
0203	Oklahoma History and Civics	____	____	____
0304	Physical Science	____	____	____
0305	Biology I	____	____	____
0403	Remedial Math	____	____	____
0404	General Math	____	____	____
0405	Algebra I	____	____	____
0408	Plane Geometry	____	____	____
0503	Art I	____	____	____
0553	Instrumental 9	____	____	____
0563	Vocal Music 9	____	____	____
0577	Orchestra	____	____	____
0703	Spanish I	____	____	____
0726	Latin I	____	____	____
0803	Home Economics I	____	____	____
0903	Woodworking I	____	____	____
0909	Mechanical Drawing I	____	____	____
1103	Physical Education 9	____	____	____
1107	Physical Education Assistant	____	____	____
1506	Library Assistant	____	____	____

Figure 12.5. Ninth-Grade Transcript Form

TRANSCRIPT II - GRADES 10 - 12

ID NUMBER_ _ _ _ _ _ NAME:______________________________
(do not punch)

Course Number	Course Name	Grade Level	Letter Grade First Sem	Letter Grade Second Sem	Credits
0104	English II	____	____	____	____
0105	American Literature	____	____	____	____
0106	English Literature	____	____	____	____
0107	World Literature	____	____	____	____
0109	H. English II	____	____	____	____
0123	Speech I	____	____	____	____
0124	Speech II	____	____	____	____
0127	Drama	____	____	____	____
0137	Journalism I	____	____	____	____
0138	Journalism II	____	____	____	____
0139	Library Science	____	____	____	____
0144	Yearbook	____	____	____	____
0145	Newspaper (Sooner)	____	____	____	____
0145	Pub. Lab. (Col Hi)	____	____	____	____
0146	Sp. English II	____	____	____	____
0149	English III	____	____	____	____
0150	Humanities	____	____	____	____
0160	Grammar & Comp.	____	____	____	____
0162	Comp/Psych	____	____	____	____
0167	Psych/Comp.	____	____	____	____

Figure 12.6. Grades 10 Through 12 Transcript Form

Figure 12.6. (Continued)

Course Number	Course Name	Grade Level	Letter Grade First Sem	Letter Grade Second Sem	Credits
0205	American History				
0206	World History				
0208	Economics				
0209	Government				
0210	Psychology				
0211	Psych/Soc				
0212	Conservation				
0213	Sociology				
0215	Prob. of Democracy				
0218	World of Geography				
0220	Ancient History				
0228	Modern History				
0230	Anthropology				
0240	Contemp. History				
0250	Consumer Economics				
0251	Soc/Psych				
0260	Economics/Government				
0305	Biology I				
0306	Biology II				
0307	Physics I				
0308	Physics II				
0309	Chemistry I				
0310	Chemistry II				

Figure 12.6. (Continued)

Course Number	Course Name	Grade Level	Letter Grade First Sem	Letter Grade Second Sem	Credits
0320	Basic Science				
0326	Earth Science				
0330	Aerosp. Educ.(Soph)				
0331	Aerosp. Educ.(Jr)				
0332	Aerospace Education				
0404	Basic Math				
0405	Algebra I				
0406	Algebra II				
0407	Sp. Algebra II				
0408	Plane Geometry				
0409	Sp. Plane Geometry				
0414	Math IV				
0416	Math V				
0424	Sp. Math IV				
0427	Sp. Math V				
0503	Art I				
0504	Art II				
0505	Art III				
0506	Art IV				
0510	Commercial Art				
0553	Band				
0563	Girls Glee Club				
0564	Mixed Chorus				

Figure 12.6. (Continued)

Course Number	Course Name	Grade Level	Letter Grade First Sem	Letter Grade Second Sem	Credits
0566	Music Theory				
0567	Concert Choir				
0577	Orchestra				
0603	Bookkeeping I				
0604	Bookkeeping II				
0609	Bus.Eng./Bus. Arith.				
0608	Introduction to Bus.				
0610	Office Practice				
0614	Shorthand				
0617	Typing				
0703	Spanish I				
0704	Spanish II				
0705	Spanish III				
0713	French I				
0714	French II				
0715	French III				
0720	German I				
0723	German II				
0724	German III				
0727	Latin II				
0728	Latin III				
0729	Latin IV				
0730	Latin III - IV				

Figure 12.6 (Continued)

Course Number	Course Name	Grade Level	Letter Grade First Sem	Letter Grade Second Sem	Credits
0903	Woodwork I				
0904	Woodwork II				
0905	Woodwork III				
0906	Machine Shop				
0909	Mech. Drawing I				
0910	Mech. Drawing II				
0917	Power Mechanics I				
0918	Power Mechanics II				
0919	Power Mechanics III				
0920	Welding				
0933	Electricity				
0945	Appliance Repair				
1007	Home Economics I				
1008	Home Economics II				
1009	Home Economics III				
1010	Home Economics IV				
1067	Distrib. Educ. I				
1068	Distrib. Educ. II				
1070	Industr.Coop. Train.				
1077	Trade Carpentry				
1095	COE (2 & 4th hr.)				
1096	COE (3 & 4th hr.)				
1097	Family Living				

Figure 12.6 (Continued)

Course Number	Course Name	Grade Level	Letter Grade First Sem	Letter Grade Second Sem	Credits
1103	P. E., Boys				
1104	P. E., Girls				
1105	P. E., Girls (adv)				
1106	P. E., Boys (adv)				
1107	Phys. Ed. Assistant				
1113	Team Phys. Ed.				
1105	Driver Education				
1090	Air Cond. & Refrig. I				
1091	Air Cond. & Refrig.				
1016	Auto Body Repair I				
1017	Auto Body Repair II				
1013	Auto Mechanics I				
1014	Auto Mechanics II				
1092	Building Trades				
1087	Data Processing				
1085	Drafting I				
1086	Drafting II				
1029	Elec./Electronics I				
1030	Elec./Electronics II				
1096	Health Services				
1037	Machine Shop I				
1038	Machine Shop II				
1064	Welding I				

Figure 12.6. (Continued)

Course Number	Course Name	Grade Level	Letter Grade First Sem	Letter Grade Second Sem	Credits
1065	Welding II				
1088	Computer Programming				
1089	Mid Management				
1098	Practical Nursing				
1210	Computational Skills (Spec. Ed 10th gr)				
1211	Communication Skills (Spec. Ed 10th gr)				
1213	Life Science (Spec. Ed 10th gr)				
1214	Social Studies (Spec. Ed 10th gr)				
1215	Computational Skills (Spec. Ed 11/12 gr)				
1216	Communication Skills (Spec. Ed 11/12 gr)				
1217	Social Studies				
1506	Library Assistant				
1507	Study Hall				
1508	Office Assistant				
1517	Cafeteria Employee				
1518	Lab Assistant				
1519	Typing Assistant				
1520	Work Release				
1521	I.C.T. Lab				
1522	D. E. Lab				

Figure 12.6 (Continued)

Course Number	Course Name	Grade Level	Letter Grade First Sem	Letter Grade Second Sem	Credits
1523	C.O.E. Lab	____	____	____	____
1524	Voc.Coop.Tr.Lab. (Special Ed)	____	____	____	____
1525	Stagecraft	____	____	____	____
1526	Audio Visual Asst.	____	____	____	____
1527	MP Assistant	____	____	____	____
1528	Teacher Assistant	____	____	____	____
1529	Band Assistant	____	____	____	____
1530	Release	____	____	____	____

HEALTH PROFILE

ID NUMBER_ _ _ _ _ _

NAME: ____
Last First Middle

Instructions

This form should be completed by the school nurse. She should rate you on each of the factors below by circling the appropriate number. The meanings assigned to the codes are as follows:

3 Normal

2 Subnormal

1 Very poor

To be Completed by School Nurse

ITEM	RATING		
1. EYES	1	2	3
2. EARS	1	2	3
3. TEETH	1	2	3
4. GROWTH	1	2	3
5. HEART	1	2	3
6. FEET	1	2	3
7. RESPIRATION	1	2	3
8. BLOOD	1	2	3
9. SKIN	1	2	3
10. WEIGHT	1	2	3

Figure 12.7. Health Profile Form

CURRENT SCHEDULE

ID_ _ _ _ _ _ _ NAME: ____________ ____________ ____________

Last First Middle

PERIOD	COURSE NO.	COURSE NAME	LOCATION	TEACHER NO.	TEACHER NAME
1.					
2.					
3.					
4.					
5.					
6.					
7.					

Figure 12.8 Current Schedule Form

GRADE POINT DATA

ID_ _ _ _ _ _   NAME________________ ________________ ________________
Last   First   Middle

GRADE	TOTAL EARNED CREDITS	TOTAL HONOR POINTS
7	______	______
8	______	______
Area	______	______
Jr. High	______	______

Figure 12.9. Grade Point Data Form

GENERAL UPDATE FORM

TO: DATA CONTROL

FROM: ______________________________ DATE ______________

PLEASE MAKE THE FOLLOWING CHANGE FOR:

ID _ _ _ _ _ _ NAME ____
Last First Middle

ITEM NAME * ____________________

ELEMENT NAME * ____________________

ELEMENT LENGTH * ____________________

DATA VALUE ____________________
(Left Justified)

* SEE USER'S MANUAL, PART VII.

Figure 13.1. General Update Form

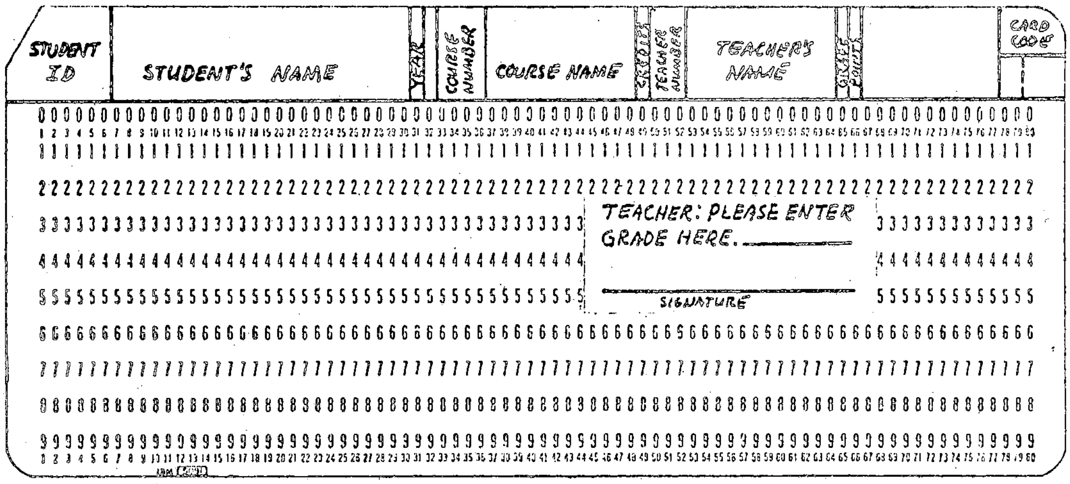


Figure 13.2. Transcript I Update Card Form

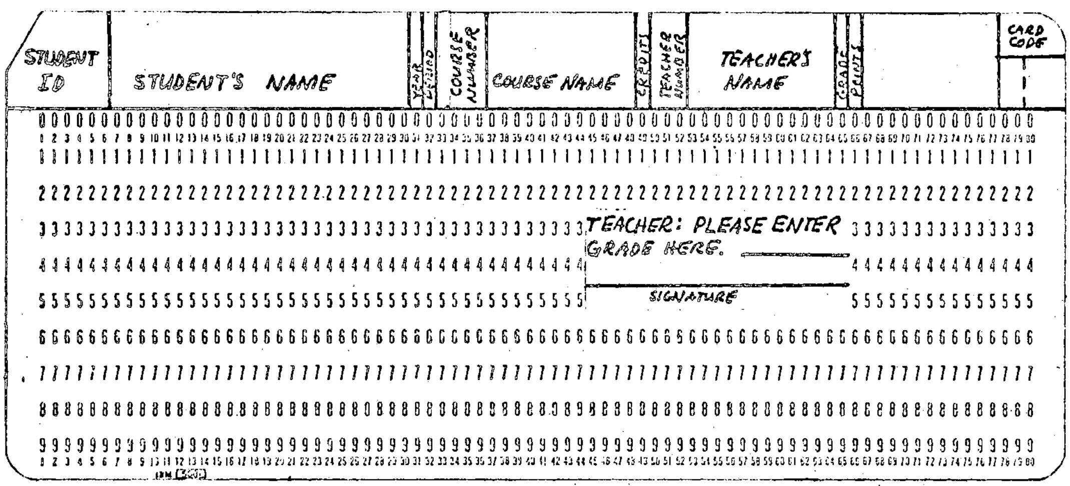


Figure 13.3. Transcript II Update Card Form

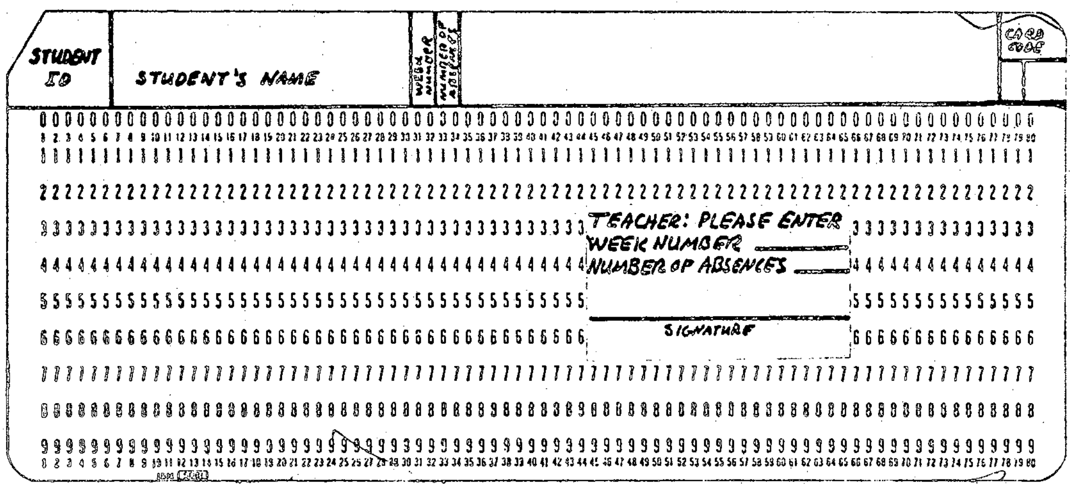


Figure 13.4. Weekly Absences Update Card Form

DELETE FORM

TO: DATA CONTROL

FROM: ______________________

(Counselor Name

PLEASE DELETE THE FOLLOWING STUDENT FROM THE COMPUTER FILES:

ID______________________ NAME ______________________________
(Last, First, Middle)

DELETION DATE ______________________

REASON _

Figure 13.5. Student Record Delete Form

STANDARDIZED TESTS UPDATE FORM

TO: DATA CONTROL

FROM: ______________________ DATE ____________

PLEASE MAKE THE FOLLOWING ( ) CHANGE ( ) ADDITION:

DC USE ONLY
TEST CODE

ID _ _ _ _ _ _ _ _ _ _ _ _ _

NAME ______________________
Last First Middle

KUHLMANN-ANDERSON _ (1)

SRA (2)

Part											
1	2	3	4	5	6	7	8	9	10	11	
12	13	Comp									

DAT (3)

Part								
1	2	3	4	5	6	7	8	9

Figure 13.6. Standarized Tests Update Form

TABLE DATA SET UPDATE FORM

TO: DATA CONTROL

FROM: ______________________________ DATE __________
Counselor Name

PLEASE MAKE THE FOLLOWING ( ) CHANGES ( ) ADDITIONS
( ) DELETION

FILE NAME (CHECK ONE): ( ) COURSE-220 ( ) DISEASE-250
( ) TEACHER-230 ( ) CHURCH -260
( ) DOCTOR-240

	CODE NO.	NAME	PHONE
1.	______	______________________________	______
2.	______	______________________________	______
3.	______	______________________________	______
4.	______	______________________________	______
5.	______	______________________________	______

* DOCTOR'S ONLY

Figure 13.7. Table Data Set Update Form

VITA

Edwin Forsberg

Candidate for the Degree of

Doctor of Education

Thesis: A DATA BASE DESIGN FOR A SECONDARY SCHOOL GUIDANCE INFORMATION SUPPORT SYSTEM

Major Field: Higher Education

Biographical:

Personal Data: Born in St. Paul, Minnesota, May 17, 1929, the son of John E. and Anna M. Forsberg.

Education: Attended grade school in St. Paul, Minnesota; graduated from Roberts Union High School, Roberts, Wisconsin, in 1947; received a Bachelor of Science in electronic engineering from the Milwaukee School of Engineering, Milwaukee, Wisconsin, in September, 1957; received the Master of Business Administration from Oklahoma State University in August, 1966; completed the requirements for the Doctor of Education at the Oklahoma State University in May, 1970.

Professional Experience: Served as a technical instructor of aircraft and engine mechanics, and applied electronics while in the U. S. Air Force from February, 1948 to February, 1954; was an instructor of electrical engineering at the University of Colorado, Boulder, Colorado, from September, 1957 to May, 1958; was a development engineer for the Exploration Department, Continental Oil Company, Ponca City, Oklahoma, from May, 1958 to March, 1961; was an equipment standards engineer for the Central Computer Department, Continental Oil Company from March, 1961 to May, 1965; was the director of data processing at the Northern Oklahoma College, Tonkawa, Oklahoma, from September, 1965 to June, 1968; currently the director of computer science at Northwestern State College, Alva, Oklahoma.

Professional Organizations: Member of the Association for Computing Machinery, Oklahoma Education Association, and Phi Delta Kappa.