DOCUMENT RESUME

ED 191 131 EA 012 850

AUTHOR Anderson, David O.
TITLE Operating Policies and Procedures of Computer Data-Base Systems.
PUB DATE Feb 80
NOTE 8p.: Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA, April 7-11, 1980). Not available in paper copy due to light print of original document.

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Confidential Records: *Data Bases: *Design Requirements: Guidelines: Higher Education: *Privacy: Students

ABSTRACT

Speaking on the operating policies and procedures of computer data bases containing information on students, the author divides his remarks into three parts: content decisions, data base security, and user access. He offers nine recommended practices that should increase the data base's usefulness to the user community: (1) the cost of developing and maintaining a computer data base must be justified in terms of valid, ongoing research questions and a large student population: (2) the data base should contain only accurate information necessary for answering important well-designed questions; (3) a formal, identifiable "body" must decide on the data base's content and the student population to be included in the data base: (4) identification numbers, rather than student names should be put on each record: (5) the coding of data must be compatible with the data base structure and consistent across years; (6) accurate, up-to-date documentation of the data base contents and coding scheme must be maintained: (7) the number of persons who have direct access to the data base must be severely limited: (8) several backup copies of the data base must be maintained: and (9) a system of priorities for performing requested analyses must be developed. (Author/IRT)

OPERATING POLICIES AND PROCEDURES
OF COMPUTER DATA-BASE SYSTEMS

David O. Anderson, Ph.D.

Office of Medical Education
University of Arizona
Tucson, Arizona 85724

February 1980

To be presented as part of a Symposium on "Issues in the Design and Implementation of Computer-based Student Information Systems," at the annual meeting of the American Educational Research Association (Boston, Mass.), April 1980.

Operating Policies and Procedures of Computer Data Base Systems

David O. Anderson

Office of Medical Education
University of Arizona

The operating policies and procedures of a computerized student data-base ultimately help determine the usefulness of that data-base. Careful consideration must be given to many aspects of that data-base: Who decides the variables to be included in the system? Which variables should they be? How will that data be coded? Who will have access to the data-base for analysis purposes? Who will have access for data entry and data deletion? How will unauthorized use of the data-base be prevented? Which users will have priority? Should there be a Steering Committee? Which data should remain confidential and what means are available to insure that? Who pays for the requested analyses?

Different approaches to these important concerns are being taken by various institutions and can offer guidance to others interested in developing or improving their own student data-base system. These concerns can be addressed under the rubrics of Content Decisions, Coding Decisions, Data-Base Security and User Access.

Content Decisions

It is important that the data collected and stored in the data-base be relevant to the issue at hand. This can include demographic, attitudinal and achievement information. Without some clear-cut guidelines to regulate content, it becomes easy to fall into the trap of gathering and storing every bit of information about the student population which is available. This amount of data soon becomes unwieldy and expensive in terms of the time and effort spent collecting, coding and storing the irrelevant, unusable information. Typical data-base information would include the following:

DEMOGRAPHIC INFORMATION

Student name (not recommended)
Identification number
Sex
Date of birth
Marital status
Fathers occupation code
Undergraduate degree level
Undergraduate school code
Undergraduate major
Summer employment codes
Extracurricular activity codes
County of residence
Year of Matriculation into the program

ACHIEVEMENT INFORMATION

Undergraduate GPA's
Admission's Test subscores
Class rank
Exam scores for each course/grades
National/State certification test scores
Date of graduation from the program
Sequence of courses through the program

ATTITUDINAL INFORMATION

Orientation Day expectations inventory responses
Career attitude inventory responses
Instructional ratings of courses/instructors
Graduation-exit questionnaire responses
Post-graduation questionnaire responses

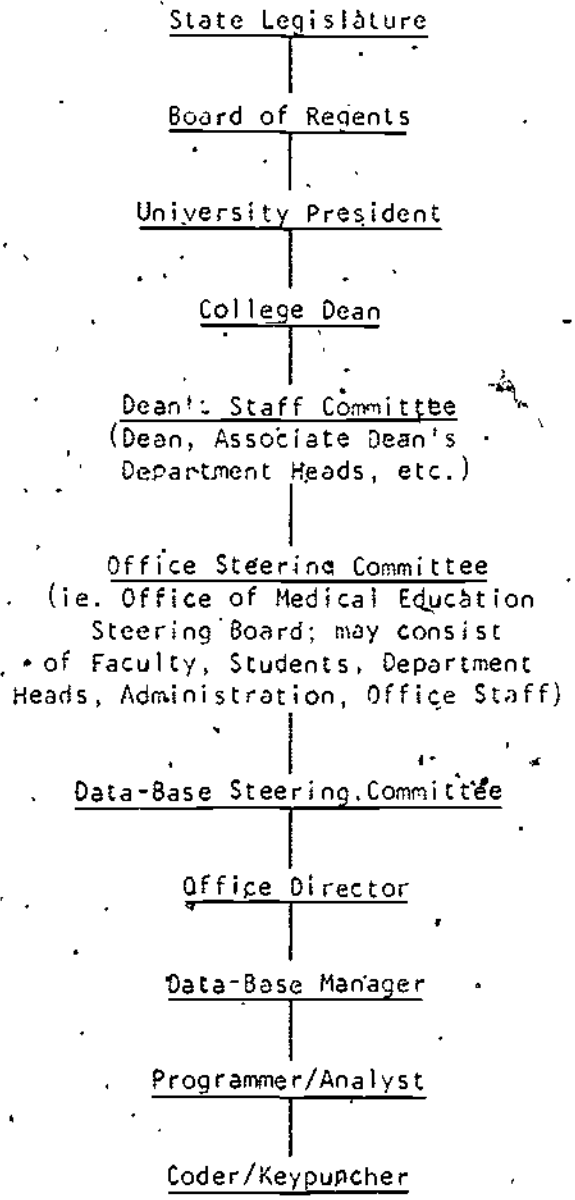
These content decisions must of course, be in agreement with all pertinent federal and state legislation governing individual privacy and informed consent.

Many computer data-bases, for the sake of confidentiality, do not contain actual student names, only student identification numbers. The correspondence between names and numbers is then kept in written form in a notebook located at a separate, secure location. Two lists of name/number pairing should be maintained: first sorted by number, then sorted by name.

Not only must the specific types of information to be stored be decided, but also the range of the student population must be considered. Is the population to be only the currently-enrolled students or all students ever enrolled in the program, even for a short period of time? If the program has a long history, perhaps data coding should start with the current class, or some specified earlier class.

These types of content decisions are of vital importance in determining the scope and ultimate utility of the data-base. Maintaining the correct data on the correct population for each study is maximally cost-effective. Going back several times through written student folders to collect those additional pieces of information not previously thought important is slow and costly.

Because these content decisions are so important, a clearly defined "body" must have that authority. That "body" can range from the University President to the data key-puncher. The figure below shows a hierarchy of possible decision-makers. Most educational offices maintaining a data-base have some sort of steering committee overseeing its operations and setting its policy and objectives. One such medical education office has a Data-Base Steering Board consisting of 3 faculty members, the Associate Dean, the Associate Registrar, and the Director of the data-base office. Another medical education office has no steering committee, leaving the office director and data-base manager solely responsible for content decisions. The decision-making "body" should be as broad as the user population (those people planning to use the data-base information).



State Legislature

Board of Regents

University President

College Dean

Dean's Staff Committee
(Dean, Associate Dean's
Department Heads, etc.)

Office Steering Committee
(ie. Office of Medical Education
Steering Board; may consist
of Faculty, Students, Department
Heads, Administration, Office Staff)

Data-Base Steering Committee

Office Director

Data-Base Manager

Programmer/Analyst

Coder/Keypuncher

Coding Decisions

Once decisions have been made as to the specific data to be incorporated into the computer data-base, its coding must be considered. The coding format will be dependent upon the organizational structure of the data-base itself. This structure can range from simply punched data cards stored in a card file, to cards read "as-is" onto magnetic tape, to highly complex data-base organizations specific to the computer hardware in use at that institution. Because the data-base manager and programmer/analyst are most familiar with the data-base structure, they should decide on the appropriate coding format.

Whatever the data-base structure, there are two requirements common to all. First, each record must have some identification field containing a unique number associating that data record with a specific individual. Secondly, if an individual has several records, these records must be numbered uniquely. That is to say, each data record must have at least an identification number referring to a specific student, and may, if necessary, have a record number.

(Student ID)	(Record No.)	(Specified Information)
10001	01	- - - - - - - - - - - - - -
10001	02	- - - - - - - - - - - - - -
10003	01	- - - - - - - - - - - - - -
11005	01	- - - - - - - - - - - - - -
11005	02	- - - - - - - - - - - - - -

Most data is coded and punched on 80-byte data cards, although, for some high-volume projects, data inout is done directly through computer terminals.

Variables to consider when deciding about coding formats include: the type of data, the frequency with which data will be collected on a regular basis, and the intended use of the data. If student names are entered into the data base, allow a sufficient number of columns for the whole name. As married students take on hyphenated last names, this field-width requirement is expanding (40 columns is sufficient).

If the same data will be coded on regular occasions (say on subsequent classes of students) the watchword is CONSISTENCY. The same type of data should be coded in the same columns across occasions. In many medical education data-bases this is not the case. Because new, revised questionnaires are used with each graduating class, the same question and associated answers are located at different spots on the questionnaire and associated punched card. It becomes a masterful task indeed, to combine several classes of data when the relevant data is sometimes in column 6, or 8, or 64! Careful pre-planning of data collection instruments and coding formats will prevent (or at least minimize) this possibility.

In addition, it is imperative that accurate, up-dated documentation be maintained of the data-base contents and the coding scheme used for each record. Too much documentation is never a problem. Keep a file of all data collection instruments, coding/keypunching instructions and a printed listing of the data-base. Being unable to locate a given piece of data which is known to be on the data-base is very frustrating; likewise, being unable to identify a piece of data you do locate can be disconcerting.

Given the opportunity to help design the data collection instrument (i.e., questionnaire), provide thoughtful concern to the ease of coding or direct keypunching. If the form is designed correctly, coding can be eliminated and data can be directly keypunched, thus saving both expense and a possible source of inaccurate data.

All data coded and keypunched should be verified for accuracy. It is much easier to correct inaccurate data at time of entry rather than after it has been included in the data base. For additional accuracy, have another person compare each line of the coding sheets to the original data, and then compare the punched cards with the coding sheets. Lastly, insure that all cards are actually and correctly entered into the data-base itself, by listing out the data-base contents for verification.

For those who may be receiving punched data cards from sources outside your own office for input to the data-base, pre-consultation is a necessary but not sufficient condition to insure compatability between the incoming data cards and the data base. However, the programmer/analyst can "easily" knock out a small reformating/recoding computer program to handle the non-standard format!!

Data-Base Security

In order for the information contained in the data-base to remain confidential and under your control, it must be "secured". This can be handled through front-end password requirements and limited user access to the data-base. This limited access strategy will be discussed in the next section of this paper.

Typically, a university computer center requires input of a secret billing number and password to directly access the computer system. This is the most general (and most easily circumvented) security layer. Finer layers include additional passwords to access data-base, to read portions of the data-base file, and finally to write or delete records. It seems obvious that these passwords should be known to only selected personnel with a true "need to know". These passwords should be changed periodically to decrease the chance of a "leak". However, be sure the passwords are recorded for reference in the event of faulty memory or the death of the only person who knows them.

For those ultra-security-conscious persons, some computer systems have the capacity to scramble the data to make it "non-sense" to the casual onlooker. Again, be able to unscramble it for analysis purposes.

Maintain several back up copies of your data-base. It is typical to have two current versions on magnetic tape stored at the computer center, perhaps an older version stored there also, and at least one current version stored at another secure location. There have been fires at computer centers with resultant losses of computer tapes! If punched cards are used to enter data into your data-base, these can serve as a suitable backup source in most circumstances, but should be stored elsewhere, preferably off-campus.

Another aspect of data-base security, often overlooked, is the trustworthiness of the data coder and keypuncher. The data coder (and sometimes the keypuncher) will be handling confidential, very identifiable information. They must be warned of its confidential nature. Indeed, if you farm out your keypunching, make sure the data are anonymous, identified by coded ID only.

User Access

Closely related to data-base security is the need to limit user access. An important distinction must be made between the handlers of the data-base itself and the users of data-base information. Only the data-base manager and perhaps the programmer/analyst should have direct access to the data-base. Other persons should have to go through these two people to have their requests for analyses fulfilled. The data-base staff will retrieve the appropriate data, perform the required analyses and then return the printed results or typed reports to the person requesting the analysis.

Because of the sensitive nature of some of the data, potential users of the data-base information should be screened, and such requests should be prioritized. The same "body" that decided on content often sets these priorities and limitations.

Requests can come from several routes, but all should result in a printed request form. This form should be initialed to indicate (1) approval for analysis, (2) completion of analysis, and (3) when the results are returned to the requestor; and (4) eventually filed.

Cost recovery: who will pay for these analyses? At some offices, legitimate requests are handled "free", the total data-base office budget coming solely from the college. In other offices, departments and funded projects pay all associated requested analysis costs, incrementing the budget of the data-base office. This policy will probably be decided at the Steering Committee level or a higher campus level.

Conclusions

Although no two computer data-bases will have exactly the same environment, purposes, format or content, it remains possible to list several recommended practices which will increase its usefulness to the user community.

1. The cost of developing and maintaining a computer data-base must be justified in terms of valid on-going research questions and a large student population.

2. The data-base should contain only accurate information necessary for answering important well-designed questions.

3. An identifiable "body" must decide on the data-base content and the student population to be included in the data-base.

4. Do not include student names in the data-base; do use identification numbers on each record.

5. The coding of data must be compatible with the data-base structure and consistent across years.

6. Maintain accurate, up-to-date documentation of the data-base contents and coding schema.

7. Severely limit the number of persons who have direct access to the data-base.

8. Maintain several backup copies of your data-base.

9. Develop a system of priorities for performing requested analyses.

DOA:sc
OME 2/29/80