A Micro Based Tape Information System

by Martin Pawlocki
and Elizabeth Stephenson [1]
Institute for Social Science Research
University of California, Los Angeles

This paper is the third in a series describing the information systems of the ISSR Data Archive. The first two have dealt with an on-line bibliographic catalog of machine-readable data files and an in depth subject index, respectively (Stephenson & Bisom, 1986; Stephenson & Pawlocki, 1987). This paper deals with the management of administrative records of the contents of magnetic tapes using a database management system, dBase III. We will outline the decision-making process in the system design phase, the technical design of the system, data entry, system end products, and our future plans for augmentation of the system. We will present details of the system as it would be used by a small number of archive staff with a small to medium sized collection of materials. We will also outline the inter-relationship between this tape management system and other records and information maintained by the Data Archive.

Overview of archive information systems

The ISSR Data Archive was established in 1977, and since that time a variety of steps have been taken to make the retrieval and access of archive materials faster, easier, and more accurate. Our ability to achieve such a goal has improved dramatically with the acquisition of our IBM PC/AT. The software available and the relative inexpensiveness of storage and retrieval of information using the PC has made all the difference. The most important result has been that we are able to work within a tiered, or layered, environment and approach to information management. That is, we have been able to use the computing resources most appropriate to our own needs as well as those of the users. The Archive now maintains several types of on-line systems, accessible through different computing centers and facilities.

For the entire campus, the University Research Library (URL) maintains an on-line technical processing system called ORION, which can be used as an on-line library catalog of holdings. Within ORION, the Archive maintains its own database containing bibliographic details about studies in the Archive collection. This can be used as a browsing aid, or for locating specific titles.

This on-line catalog has some limitations for the description of machine-readable survey data.

[1] Presented at the International Association for Social Science Information Service and Technology (IASSIST) Conference held in Washington, D.C., May 26-29, 1988

The bibliographic entries cannot provide the abstracts necessary to describe in detail the content of files. Many surveys cover more topics than can be addressed by subject classification. Also, the on-line catalog is maintained on a mainframe computer, although subfiles may be downloaded to a personal computer-based database management system (DBMS).

To provide users with more detail about the content of individual data files, the Archive maintains subject oriented indexes which focus on specific areas (e.g., women's studies, ethnicity, health statistics), and contain abstracts and detailed indexing of the studies they include. Files are assigned up to 20 index terms appropriate to the subject area. The indexes are maintained for on-line searching within the Archive and will soon be available as part of a campus local area network, accessible to all users. Printed copies of the subject index are also available.

The library catalog and subject index are helpful only in identifying potentially useful data. Users of MRDF must examine variable lists and questionnaires to determine specific details about a data file. To address this need, researchers have access to the ICPSR's variable-based search facility operating under SPIRES on CDNet. In order to provide similar variable-level access to the contents of locally produced files, the Archive has developed a database containing the question text, variable names, and value codes for each question in a survey. At present, this database contains only the surveys conducted as part of the Los Angeles Metropolitan Area Surveys (LAMAS) or the Southern California Social Survey (SCSS).

When a file has been identified as being useful for research, a user needs technical information regarding the physical format and structure of the file, and its location on magnetic tape. For some time, the Archive maintained and provided this information on paper. With the acquisition of the microcomputer, and the availability of a DBMS, we began to explore the possibility of maintaining what was becoming a rather voluminous and unwieldy information system on computer. The following is a discussion of the planning, design and implementation of a tape information management system.

Preplanning and system design

Our first step, of course, was to discuss what we'd like in a system. This process took about three months of intensive effort, although we had discussed it previously as part of an ideal design for archive information management. It required that each staff member who was to use the system scrutinize every aspect of a variety of tasks, and the type of product or outcome to be produced. The staff using the system work in different areas of the archive, and with different sets of tasks and archival materials. We focused on many aspects of administrative records, reference needs, technical needs, records management for tapes, links to a campus information network, as well as the information needs of users.

The archivist viewed the system as potentially satisfying administrative and reference needs. There was a need for quick retrieval of technical and bibliographic information about every study held by the archive. The system would have to produce some paper products containing information about studies. Further, there was a need to be able to verify our holdings of particular studies by having a machine-readable shelf list.

Some of this information is traditionally stored in library catalogs. As previously mentioned, the archive does have a machine-readable catalog using UCLA's on-line system, ORION. This catalog has proven itself useful for reference work, but the MARC-type record does not provide for the technical information that it is necessary to maintain about each file. This is especially true for multi-file studies, such as those which have have up to several hundred associated data sets. There is no mechanism in ORION with which to document all these files, and there appeared to be no way to combine bibliographic and technical information about several studies for users to access. We knew that a significant number of users knew which data file they wanted to use, and needed only the tape and file details in order to begin research projects. These needs would be better satisfied with a database management system.

The programmer's mission was to design a system that could be made accessible within UCLA's token ring network, SSCNet. This network uses Novell software (Advanced Netware) and an IBM Token Ring design. We wanted to make it possible for users to search the archive's tape system from anywhere on campus. We also wanted the system design to reflect the way in which we anticipated users would query the system, so that it could be self documenting. In addition, we wanted the programmer to design a system so that others familiar with dBase III might be able to interpret system bugs and understand the overall structure of the system. The programmer's design decisions are further outlined in the technical portion of this paper.

The technical assistant was to use the system from a task oriented viewpoint. Her needs focused on ease of entering, modifying, adding, and deleting information. As she would use the system repeatedly for the same procedures, it was important that she understand many details about the system and its limitations. The technical assistant would use the system to download the DCB information about each file of a study[2], and add bibliographic and other descriptive information, including subject terms, notes about file content, and file structure or format. She also needed the system for records management purposes. We wanted to be able to verify the contents of a single reel of tape, which might contain part of a large study, or many files of smaller studies. Other tape information needed for records management purposes included dates of file creation and tape cleaning, and the remaining space available on each tape. (In our facility, we follow a policy of filling all but 100 feet of a tape where possible. Knowing the number of unused feet of tape helps the technical assistant assign studies to archival storage on tape.)

For all of the Archive staff, there was a need for the system to be microcomputer based. Access to the mainframe computer (IBM 3090 with MVS and VM/CMS operating systems) is required for maintenance and copying of tapes, but it is an expensive medium for storage of administrative information. It is difficult to use the mainframe to produce the printed products we want, to link with other information systems maintained by the archive, nor is the mainframe accessible to all campus users. That is, not everyone on the campus has the financial resources to access the mainframe.

We also wanted to use the types of software that are available only in a microcomputer environment. While there are database management systems for mainframes, such as SPIRES, FOCUS, RAMIS, they are neither widely understood nor used by campus researchers. We wanted to avoid having to train users in understanding the structure and software of a DBMS. We felt that this would be more likely avoided in a microcomputer

[2] Editor's note DCB or data control block information is SAS nomenclature which includes record (or block) format (RECFM), physical record length (LRECL), and block size (BLKSIZE).

based situation. Microcomputers are also more attractive since the campus is being linked as a network via PC's, with the concommitant potential of making this system available to the largest number of users.

The following important information components were identified in the initial phase of system design: study title, principal investigator names, subject headings, study or accession number, tape volume id number, tape file numbers for each study, DCB information for each file, notes on individual file structure and format. We considered these to be the intellectual items we would need to produce a variety of system end products.

End products and system searching

The desired end products fell into three categories: paper, on-line internal, and on-line external (or system end products). The paper product is called a Tape Information Sheet. (An example is found in the Appendix). These have been part of the record keeping and information dissemination functions of the archive since it was established. The tape information sheet contains study title, names of principal investigators, major subject heading, tape volume id number, format of the tape (density, mode, parity, labeled/non-labeled), and the file numbers, dataset names and DCB information for each file associated with the study title. The purpose of the tape information sheet is threefold: 1) maintain a bibliographic shelflist; 2) maintain a tape number shelf list; 3) provide users with printed details about studies. A copy of the tape information sheet is maintained as part of a shelflist, filed by broad subject category for each study in the Archive collection, and also stored by tape volume id number so that we can verify the content of specific tapes.

We maintain files describing the content of each magnetic tape. As new studies are acquired, they are copied to Archive tapes which use a consecutive numbering system. A user accessible copy is stored at the computer center with a tape id number prefix "DTA" followed by a three digit number. As tapes are numbered consecutively, we can continue to use this system until we reach 999 reels of tape. An archival backup copy of the "DTA" tape is stored in a separate location and has the same consecutive numbering preceeded by the prefix "DTB". The tape information system stores information using the "DTA" tape number which is identical to the "DTB" tape number.

The third function of the tape information sheet is to provide users with a paper copy of all technical and bibliographic details required for the use of a data file. We found this to be essential in order that users have accurate information about the data they wish to use. Some computing centers have eliminated the need for this by cataloguing datasets, but our center is not set up in this way, and as we have stated earlier, the mainframe is not accessible to all. Further, the catalogued data sets cannot be easily linked to bibliographic and records management information, without significant systems programming.

Our other end products were largely focused on how the system would search for and retrieve information, and what information would be used in-house versus what would be publicly available. The actual process is described in the technical description below, but will be discussed in general terms here.

Basically, we wanted to be able to search the system in several ways: by title, principal investigator, subject, study number and tape number. We wanted to retrieve all information that met the search criteria. That is, we wanted all editions of a study (for example, all versions of the American National Election Studies produced by ICPSR), or all years of a study with the same title (for example, Current Population Survey, March Annual Demographic File), but we did not want to have to specify, in the search, all editions or years, since we might not know that information. We also wanted to be able to view the complete list and select whichever files or studies were desired. In addition, we wanted to view a facsimile of the tape information sheet on the screen and, as described earlier, produce a print copy of the same information.

Some information and some types of searches were to be used only for internal purposes. For instance, several subject terms might be assigned to a study for use in searching, but only the broad intellectual category (e.g. mass political behavior) would appear on the printed tape information sheet. Also, users should not be able to perform searches by tape number, since they would be interested in specific files and would not need to know the content of a whole tape.

Data entry and authority lists

Once the system was designed and the programs written and debugged, we began an intensive period of data entry, beginning with the most heavily used types of files and entering all newly acquired data. Each tape is mapped or scanned using the mainframe, and this information is downloaded and reformatted. Bibliographic details are added through the use of screen templates, selection menus and prompts. The system was designed so that it would be easy to train ourselves and others in data entry, and when mistakes are made, to delete information or exit the system. The manner of downloading DCB information was a timesaving device and ensured that this information would be accurate and not subject to typing errors. The accuracy of the bibliographic information had been established using title/author and accession number authority lists. Subjects were assigned according to a pre-defined set of guidelines.

Subjects were meant to be broad and were to focus on categorization of the file rather than topical content. We also assigned geographic headings, acronyms of titles or principal investigators, and type of data descriptors such as "census". By and large, we tried to anticipate terminology that would be used by researchers when searching the system.

Future plans

This system will be very useful both for ourselves and for campus users until the collection becomes very large. At that time, the size of the database will be too large to accomodate the search pattern the system now employs. The campus is in the process of selecting a miniframe/microcomputer RDBMS. When this is acquired and put into place, we expect that we will be able to use its features to link all of the information systems we have developed, and perhaps provide linkage to the actual data from the files.

In the more immediate future, we would like to be able to permit users to download portions of their searches into personal files, and to produce a printed catalog of all our headings. We also expect to use the system with our detailed indexes and abstracts to create additional specialized indexes, but the design of such a project has not been completed.

Technical descriptions

Until recently, much of the Archive administrative work was done on a mainframe computer. Over the last three years, we have been in the process of transferring this work to microcomputers. The project on which we are reporting started life as a way of producing tape information sheets on the microcomputer. When we noted the elements that would need to be entered, we realized that an opportunity to set up a microcomputer-based system to keep track of studies and tapes had presented itself. Once all the data are loaded, the system could become a catalog of our collection. So what was originally intended as an administrative system could also be used as a reference tool.

One of the first questions to be decided was which database management package to use. The options, because of availability, were dBASE III and R:BASE System V. dBase was chosen based on an analysis done by a library school student comparing the two packages. It was chosen because of its flexibility, power, speed, wide use on campus, and because the programmer was already very familiar with its programming language.

The primary concerns in developing the software were that it be user-friendly, menu driven, and most important, self-documenting. Once these issues were addressed in the design specifications, the system was fine-tuned in order to speed it up. This fine-tuning process is still going on.

The data base can be searched by title, subject, principal investigator, or study number. A record is retrieved if the search term is found anywhere in the field being searched. There is also an exact match search, and a left-adjusted search. That is, the search term must match starting at the leftmost character of the field being searched. For all types of searches, the results are displayed alphabetically by title. If the search is by principal investigator or subject, and more than one subject or principal investigator satisfies the search criteria, a list of results is displayed before individual titles are shown. The search can then be refined, or continued with the original search terms. The output from searching the database is the aforementioned tape information sheet.

The TAPES system is a relational database consisting of eight files, containing four types of fields: title, principal investigators, subjects, and the information relating to the computer files on tape (see the Appendix).

The title may contain up to 256 characters. Leading English articles are bypassed when the title sort key is created. The title sort key is a five digit code which, when accessed in ascending order, will keep the titles in alphabetical order. The codes were established with gaps to facilitate the insertion of new records. When the gaps are filled and the system slows down (a subjective judgement to be sure), a utility program is run to renumber the title keys. When stored in the file, titles are broken into lines of 71 characters to speed up the display of long titles. One of the principal investigator files contains the full text form of the principal investigator and a five digit code. The code is a consecutive accession number created when a new principal investigator (i.e. one not already in the file) is added. Each name appears only once in this file. Another principal investigator file contains the principal investigator code and the title sort key of the study to which this principal investigator is attached. As the system is presently configured, there can be up to nine principal investigators per study.

There is a similar arrangement for subjects. There are two files: one consisting of the full text form of the subject and a code, and the other containing the code and the title sort key associated with the study. As with principal investigator, up to nine subjects may be assigned per study.

Three files are used to keep track of the tape file information. The primary one contains data control block (DCB) information, the number of blocks and feet, (this information is downloaded from the mainframe), study number and a code for file format and file description (added manually). A second file contains information about the entire tape. Included are the number of files on the tape, the date the information was loaded into the system, the date of the last time the tape was cleaned, amount of space remaining on the tape, and whether more files can be stored on the tape. The third file consists of a note (if nccessary) about the tape file. For example, the note field might contain the name of a state, or a description of the file.

When using the TAPES system as a means of accessing information about studies, the primary access point is the study number. The study number can be either the ICPSR number or a locally assigned number. For this later category, the study number can represent either an individual study or a group of studies. For example, instead of having each year of the March Current Population Survey (which we want to identify collectively) entered under its own individual ICPSR study number, a group number was assigned.

When used to get information about tapes, the primary access point is what is referred to in computer terminology as the tape volume serial number. Our tapes all have tape volumes starting with the letters "DTA" and numbered consecutively with a three digit number.

As we use the TAPES system, it is inevitable that problems will arise. The first modifications will be to fine-tune and correct any errors. The next major enhancement, however, will be to add abstracts to the studies.

Bibliography

dBase III; user manual. Culver City, CA: Ashton-Tate, 1984.

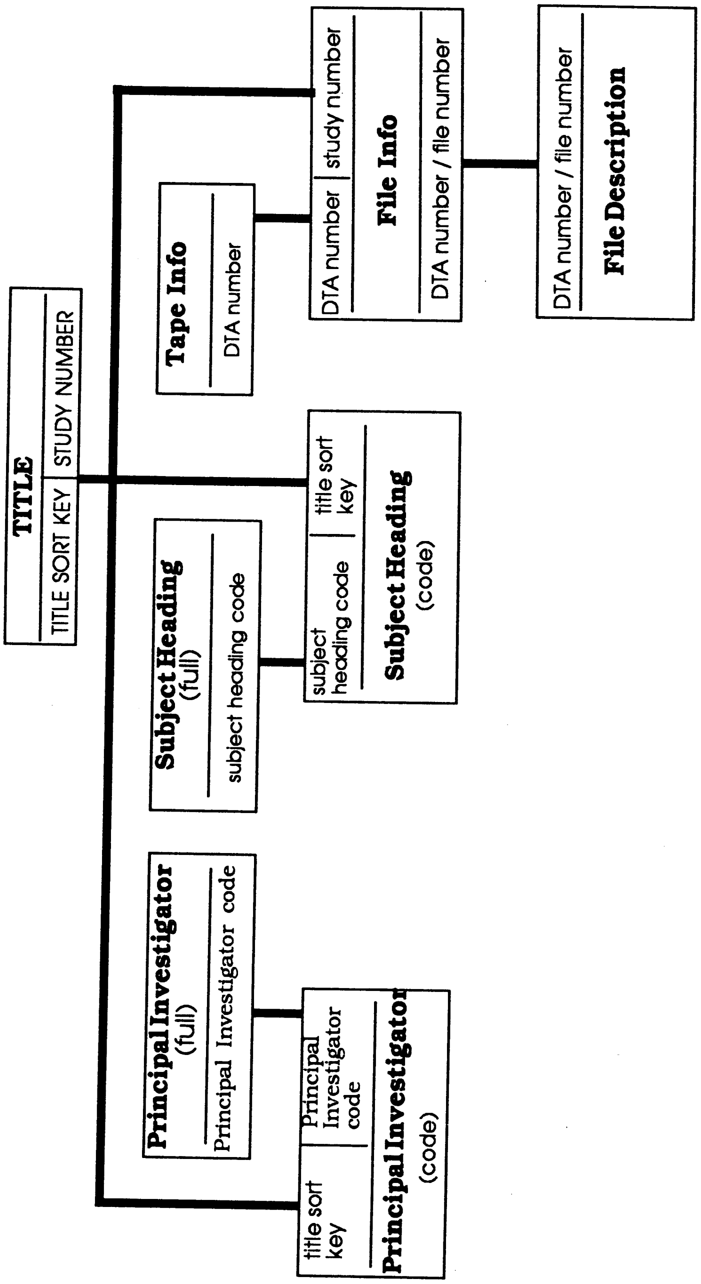
Guide to resources and services, 1987-88. Ann Arbor, MI: University of Michigan. Inter-university Consortium for Political and Social Research, 1987.

Liskin, Miriam. Advanced dBase III: programming and techniques. Berkeley, CA: Osborne McGraw-Hill, 1986.

Stephenson, Elizabeth A. and Diane Bisom. *Using ORION to CREATE an on-line catalog of data archive holdings.* Paper presented at the IASSIST Conference, Los Angeles, CA., May 21-25, 1986.

Stephenson, Elizabeth A. and Martin Pawlocki. *Index of machinereadable data files for women's studies.* Paper presented at the IASSIST Conference, Vancouver, B.C., May 19-22, 1987.

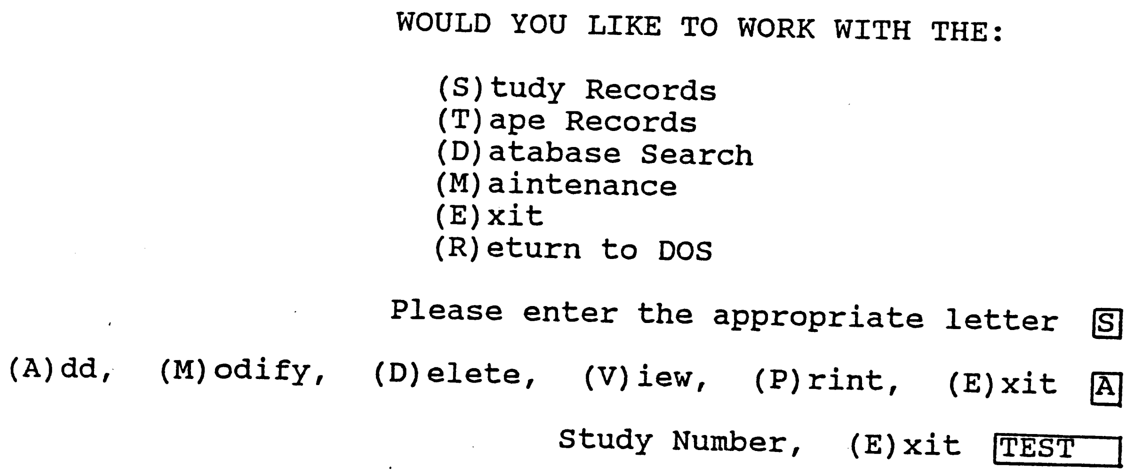
1


TITLE
TITLE SORT KEY
STUDY NUMBER
Tape Info
DTA number
DTA number
study number
File Info
DTA number / file number
DTA number / file number
File Description
Subject Heading
(full)
subject heading code
subject heading code
title sort key
Subject Heading
(code)
Principal Investigator
(full)
Principal Investigator code
title sort key
Principal Investigator code
Principal Investigator
(code)

2

```
     ***** THESE DATA ARE MADE AVAILABLE FOR USE EXCLUSIVELY *****
           ***** AT UCLA, AND MAY NOT BE REDISTRIBUTED *****

Miller, Warren E.

AMERICAN NATIONAL ELECTION STUDY, 1986

STUDY NO: I8678     VOL ID: DTA446, DTA499

TRACK: 9      DENSITY: 6250      LABEL: SL      MODE: EBCDIC      PARITY: ODD

-----------------------------------------------------------------------------
FILE        FILE                                                FILE    APPROX
DESC.       NO.      DSNAME           RECFM  LRECL  BLKSIZE  FORMAT  NO. OF
                                                                     RECORDS
-----------------------------------------------------------------------------
Dictionary   5   DI8678.CPS            FB      80     1600   OSIRIS      870
(DTA446)Post Election Survey, April 1987 (CPS version)

Data         6   DA8678.CPS            FB    1227     4908   OSIRIS     2174
(DTA446)Post Election Survey, April 1987 (CPS version)

Dictionary   7   ODICT.S8678           FB      80     1600   OSIRIS      810
(DTA499)Post Election Survey, 1986 (ICPSR Version)

Data         8   ODATA.S8678           FB    1220    29280   OSIRIS     2172
(DTA499)Post Election Survey, 1986 (ICPSR Version)

Codebook     9   OCDBK.S8678           FB      80    30000   OSIRIS    20062
(DTA499)Post Election Survey, 1986 (ICPSR Version)

Codebook     10  CB8678                FB      80    30000   LISTED    31312
(DTA499)Post Election Survey, 1986 (ICPSR Version)

Data Def.    11  SP8678.LRECL          FB      80    30000   SPSS       3937
(DTA499)Post Election Survey, 1986 (ICPSR Version)

Data         12  FQ8678                FB     132    29964   LOGICAL    2156
(DTA499)Post Election Survey Frequencies (ICPSR Version)
```

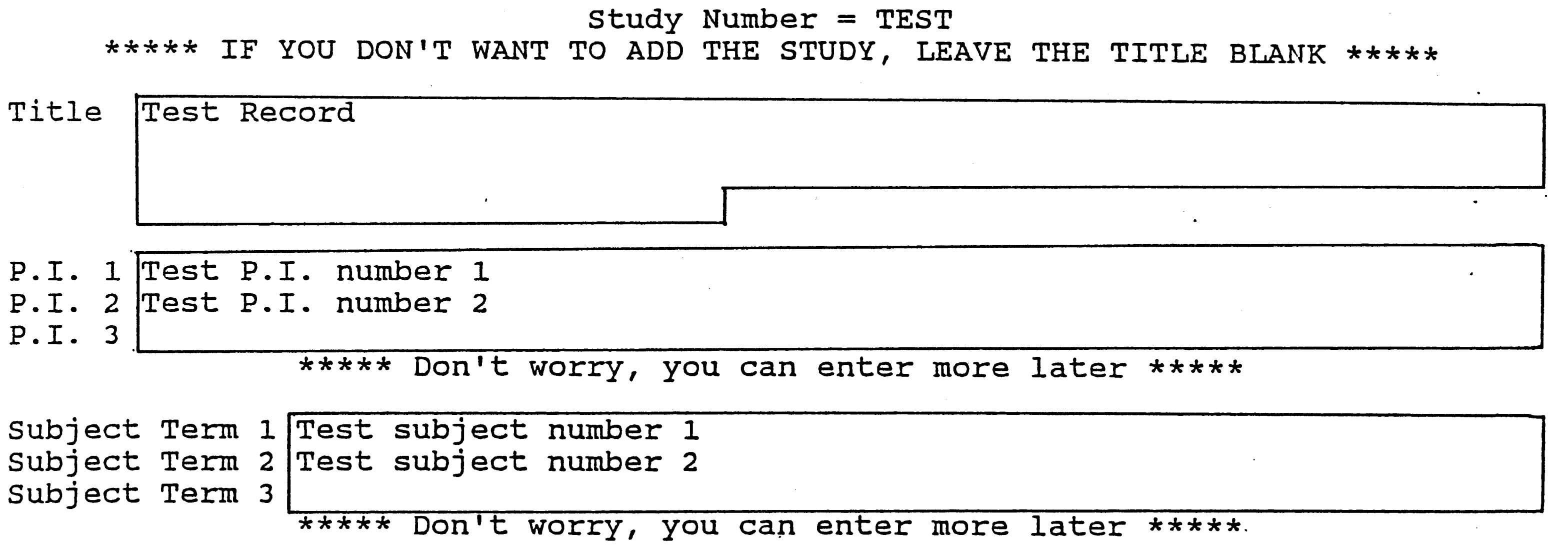
3



```
                          WOULD YOU LIKE TO WORK WITH THE:

                            (S)tudy Records
                            (T)ape Records
                            (D)atabase Search
                            (M)aintenance
                            (E)xit
                            (R)eturn to DOS

                          Please enter the appropriate letter  S

(A)dd,   (M)odify,   (D)elete,   (V)iew,   (P)rint,   (E)xit  A

                                  Study Number,   (E)xit  TEST
```

4

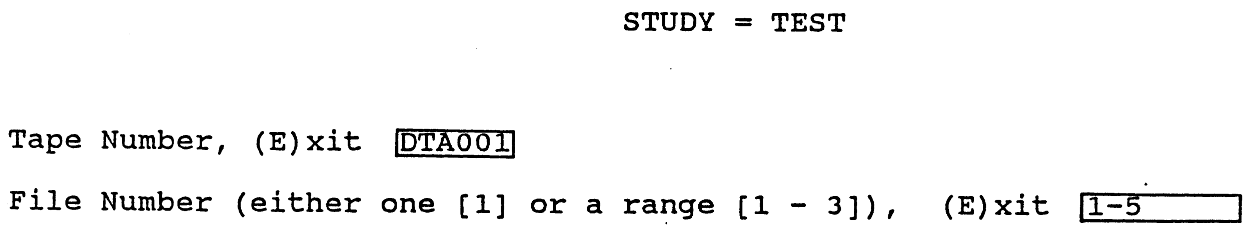
eg.4



```
                Study Number = TEST
***** IF YOU DON'T WANT TO ADD THE STUDY, LEAVE THE TITLE BLANK *****

Title   Test Record

P.I. 1 Test P.I. number 1
P.I. 2 Test P.I. number 2
P.I. 3
            ***** Don't worry, you can enter more later *****

Subject Term 1 Test subject number 1
Subject Term 2 Test subject number 2
Subject Term 3
            ***** Don't worry, you can enter more later *****
```

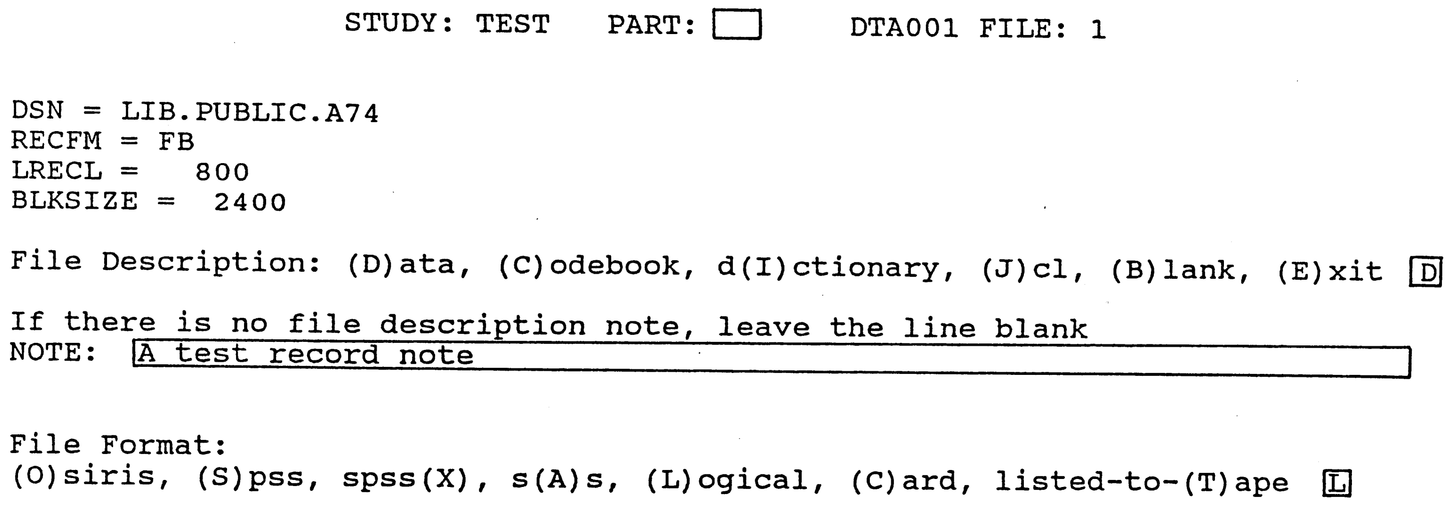
5

eg.5



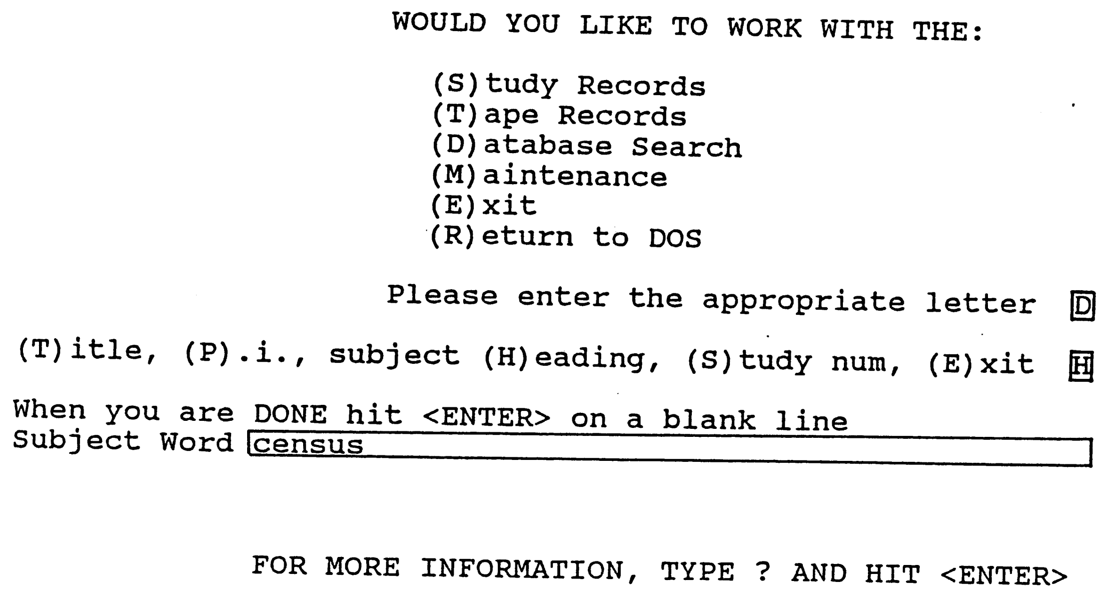
```
                    STUDY = TEST

Tape Number, (E)xit  DTA001

File Number (either one [1] or a range [1 - 3]),  (E)xit  1-5
```

6



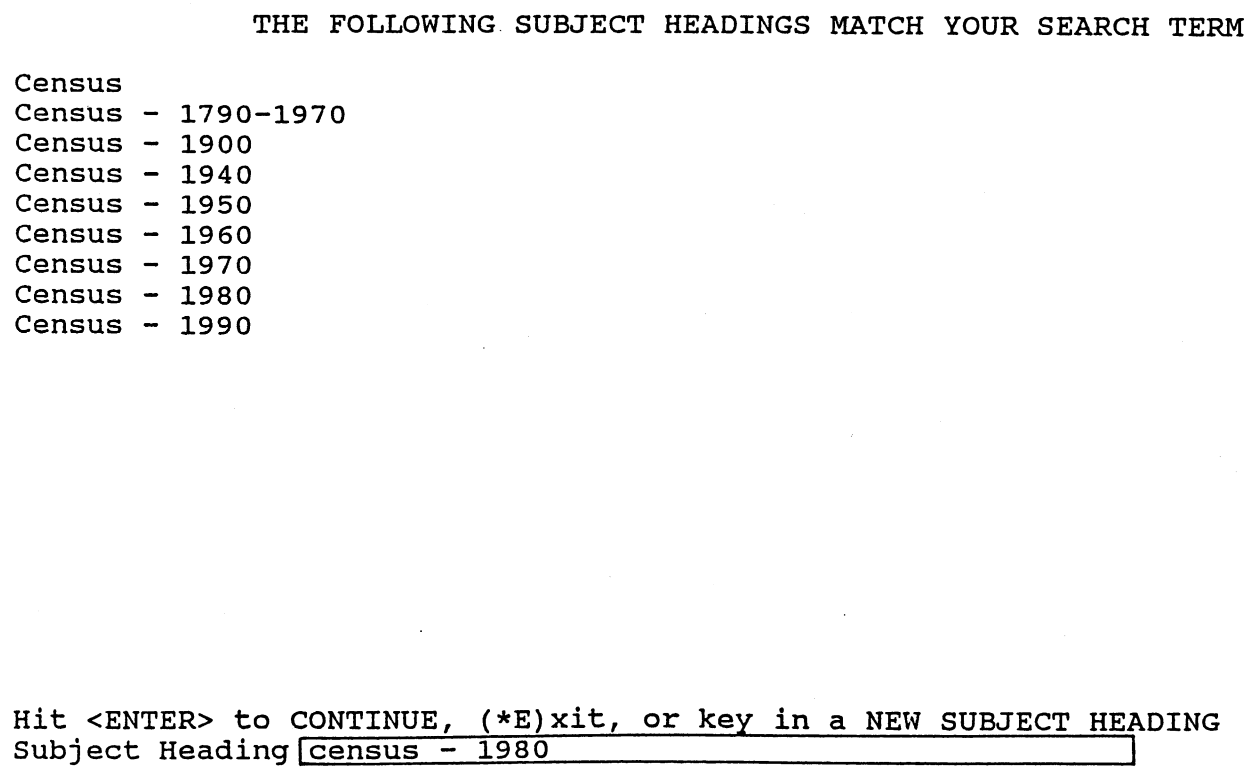
```
                   STUDY: TEST   PART: □       DTA001 FILE: 1

DSN = LIB.PUBLIC.A74
RECFM = FB
LRECL =    800
BLKSIZE =  2400

File Description: (D)ata, (C)odebook, d(I)ctionary, (J)cl, (B)lank, (E)xit D

If there is no file description note, leave the line blank
NOTE: A test record note

File Format:
(O)siris, (S)pss, spss(X), s(A)s, (L)ogical, (C)ard, listed-to-(T)ape L
```

7



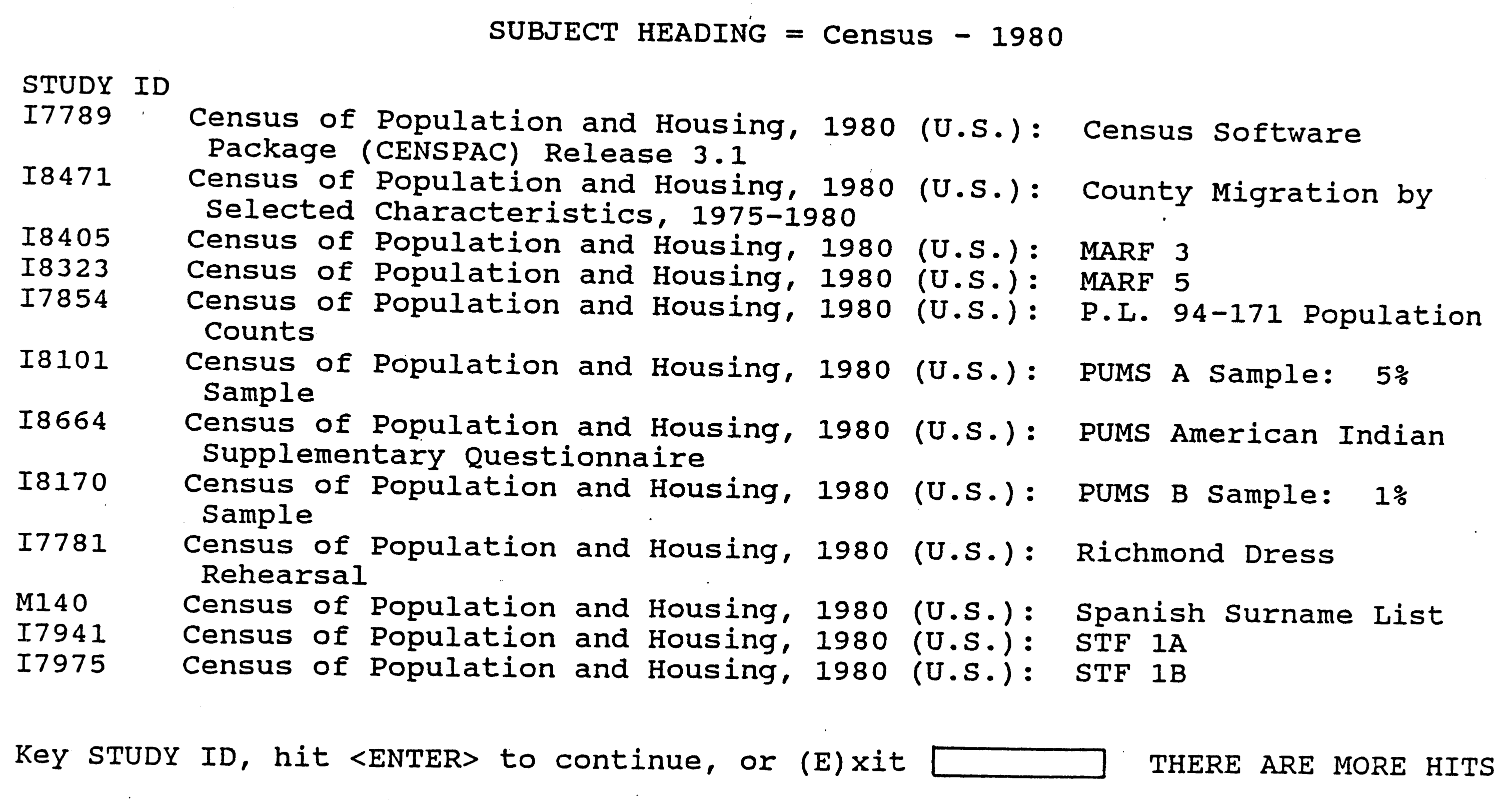
```
                WOULD YOU LIKE TO WORK WITH THE:

                  (S)tudy Records
                  (T)ape Records
                  (D)atabase Search
                  (M)aintenance
                  (E)xit
                  (R)eturn to DOS

                Please enter the appropriate letter D

(T)itle, (P).i., subject (H)eading, (S)tudy num, (E)xit H

When you are DONE hit <ENTER> on a blank line
Subject Word census


           FOR MORE INFORMATION, TYPE ? AND HIT <ENTER>
```

8



```
                 THE FOLLOWING SUBJECT HEADINGS MATCH YOUR SEARCH TERM

Census
Census - 1790-1970
Census - 1900
Census - 1940
Census - 1950
Census - 1960
Census - 1970
Census - 1980
Census - 1990




Hit <ENTER> to CONTINUE, (*E)xit, or key in a NEW SUBJECT HEADING
Subject Heading[census - 1980                          ]
```

9



```
                    SUBJECT HEADING = Census - 1980

STUDY ID
I7789     Census of Population and Housing, 1980 (U.S.):  Census Software
           Package (CENSPAC) Release 3.1
I8471     Census of Population and Housing, 1980 (U.S.):  County Migration by
           Selected Characteristics, 1975-1980
I8405     Census of Population and Housing, 1980 (U.S.):  MARF 3
I8323     Census of Population and Housing, 1980 (U.S.):  MARF 5
I7854     Census of Population and Housing, 1980 (U.S.):  P.L. 94-171 Population
           Counts
I8101     Census of Population and Housing, 1980 (U.S.):  PUMS A Sample:  5%
           Sample
I8664     Census of Population and Housing, 1980 (U.S.):  PUMS American Indian
           Supplementary Questionnaire
I8170     Census of Population and Housing, 1980 (U.S.):  PUMS B Sample:  1%
           Sample
I7781     Census of Population and Housing, 1980 (U.S.):  Richmond Dress
           Rehearsal
M140      Census of Population and Housing, 1980 (U.S.):  Spanish Surname List
I7941     Census of Population and Housing, 1980 (U.S.):  STF 1A
I7975     Census of Population and Housing, 1980 (U.S.):  STF 1B

Key STUDY ID, hit <ENTER> to continue, or (E)xit [        ]  THERE ARE MORE HITS
```