STATISTICAL INQUIRY & RETRIEVAL
an IMAGE application

MARTIN D. JEWEL
NATIONAL SPINAL CORD INJURY DATA RESEARCH CENTER

BACKGROUND

The National Spinal Cord Injury Data Research Center (NSCIDRC) - a division of Good Samaritan Hospital in Phoenix, Arizona - is supported in part by a grant from the U.S. Department of Health, Education & Welfare through the Rehabilitation Services Administration. NSCIDRC's goal is to provide access to a national repository of data relative to spinal cord injured persons for the purpose of improving the care and treatment thereof, and reducing the length of hospital stay and associated costs.

Since spinal cord injury is a sudden traumatic shock and extremely expensive, the costs are often borne by society in the form of taxes and insurance premiums. Helping a patient to achieve his most productive status as quickly as possible gives him a psychological boost, reduces the drain on family and personal resources, and decreases the cost to society.

SYSTEM FLOW

The source of patient data is the eleven Regional Model SCI Systems (see appendix 1). Data is extracted from hospital records, physicians' statements, patient interviews and bills for various types of equipment and services. This information is compiled by medical record personnel and transcribed onto pre-printed forms. The forms are assembled into batches upon completion, logged, and then forwarded to Phoenix. Generally, a batch represents a weeks work. (see Figure 1).

After receiving the batched forms at NSCIDRC, the forms are logged in on the HP3000 and sorted by new entries and updates (see Figure 2). The new entry data is keyed into the computer via an HP2645 video terminal using a general purpose data entry program designed to create a transaction file. We do not use DEL, having found it inadequate for our needs.

Processing at the Regional SCI Center

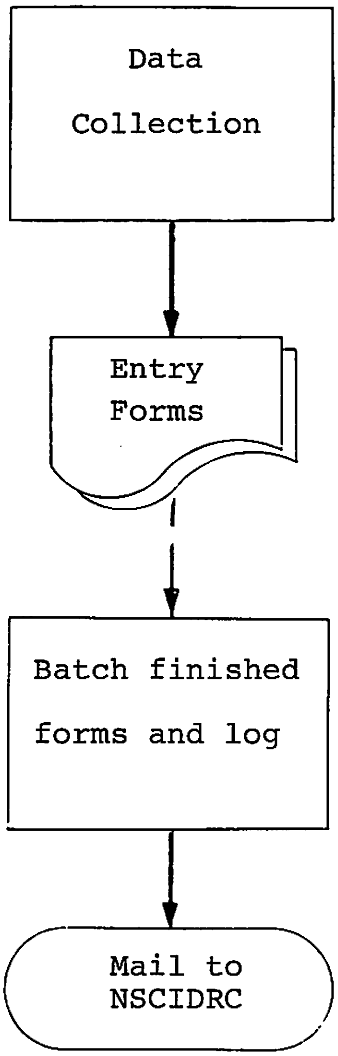


Figure 1.

Processing at NSCIDRC

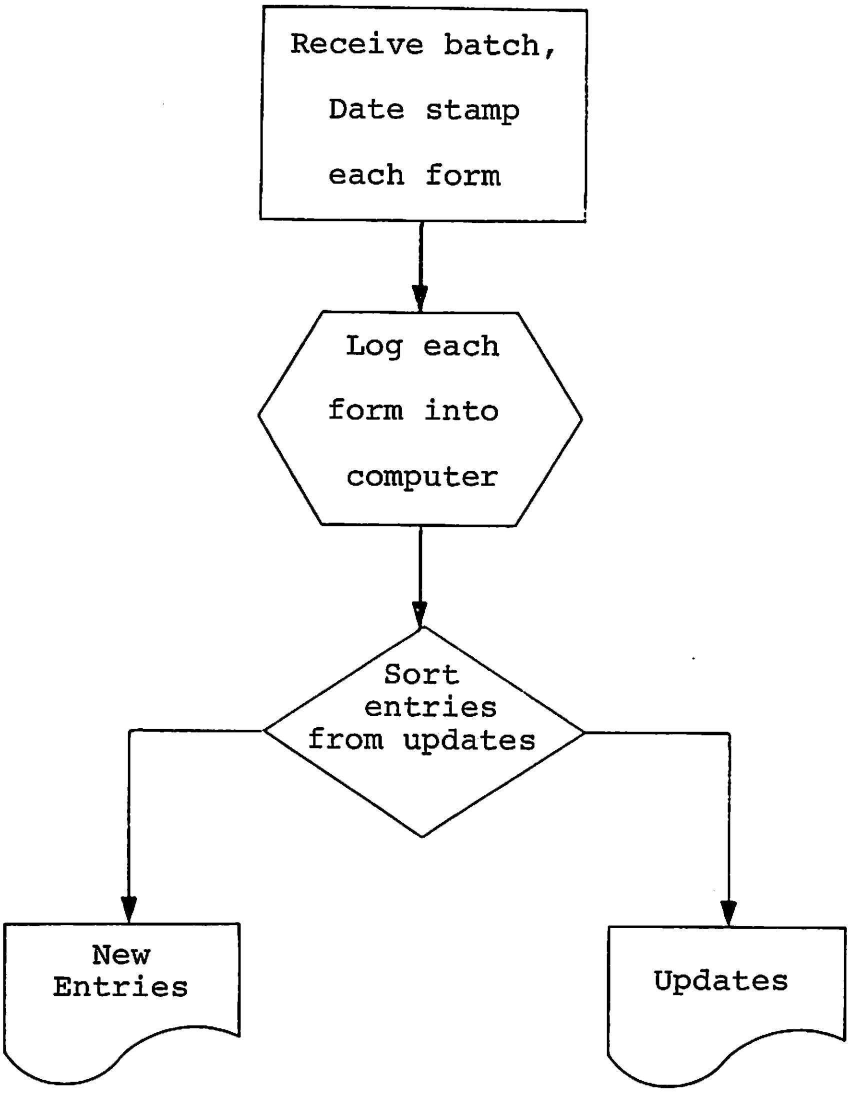


Figure 2.

At appropriate periods, usually twice a day, a data base posting program is executed as a batch job stream to add the data to the data base (see Figure 3). We do not enter data directly to the data base.

Twice monthly, a data quality audit program is run to produce a discrepancy list which is forwarded to the Regional SCI Centers for corrective measures. Resultant updates are then sent to NSCIDRC as described above. The updates are posted to the data base on-line. Data verification is done after entry/posting and updating.

Other data base management tools are shown in Figure 4. The selective dump provides a hardcopy of patient data as it exists in the data base. The Form 2 Follow-up produces a tickler report of those forms which are due during the next quarter, and an expediting list of in-process and past-due forms.

We have grown increasingly confident of the relative cleanliness of the data base. We now average fewer than 2 discrepancies per 100 forms. With the many checks for validity and logical interrelationships between variables, the error rate is approximately 1 in 40,000.

DATA FLOW TIMING

Patient information comes to NSCIDRC at the completion of the initial hospitalization and rehabilitation period, and again at each subsequent anniversary of injury. There may be a lag time of up to three months while data is extracted from case records and various professionals respond to requests for information. Typically, after three to six months after the end of a calendar year the data for that year is complete, clean and ready to be analyzed.

The form, on which the initial data is submitted, is referred to as Form 1 and is complete in itself. The annual follow-up data is reported on a Form 2. Each Form 2 may have one or more Hospitalization forms (Form H) attached, if the patient was admitted to a hospital during that year. Thus, a particular patient will have a single Form 1, and, depending on the number of years after injury, one or more Form 2's. Each Form 2 may, in turn, have one or more Form H's attached.

DATA BASE MANAGEMENT BACKGROUND

Initially, data was stored on an IBM System/32 which had many hardware and software limitations. It was limited to producing RPG reports and had no data base capabilities. In addition, the volume of data was not sufficient for

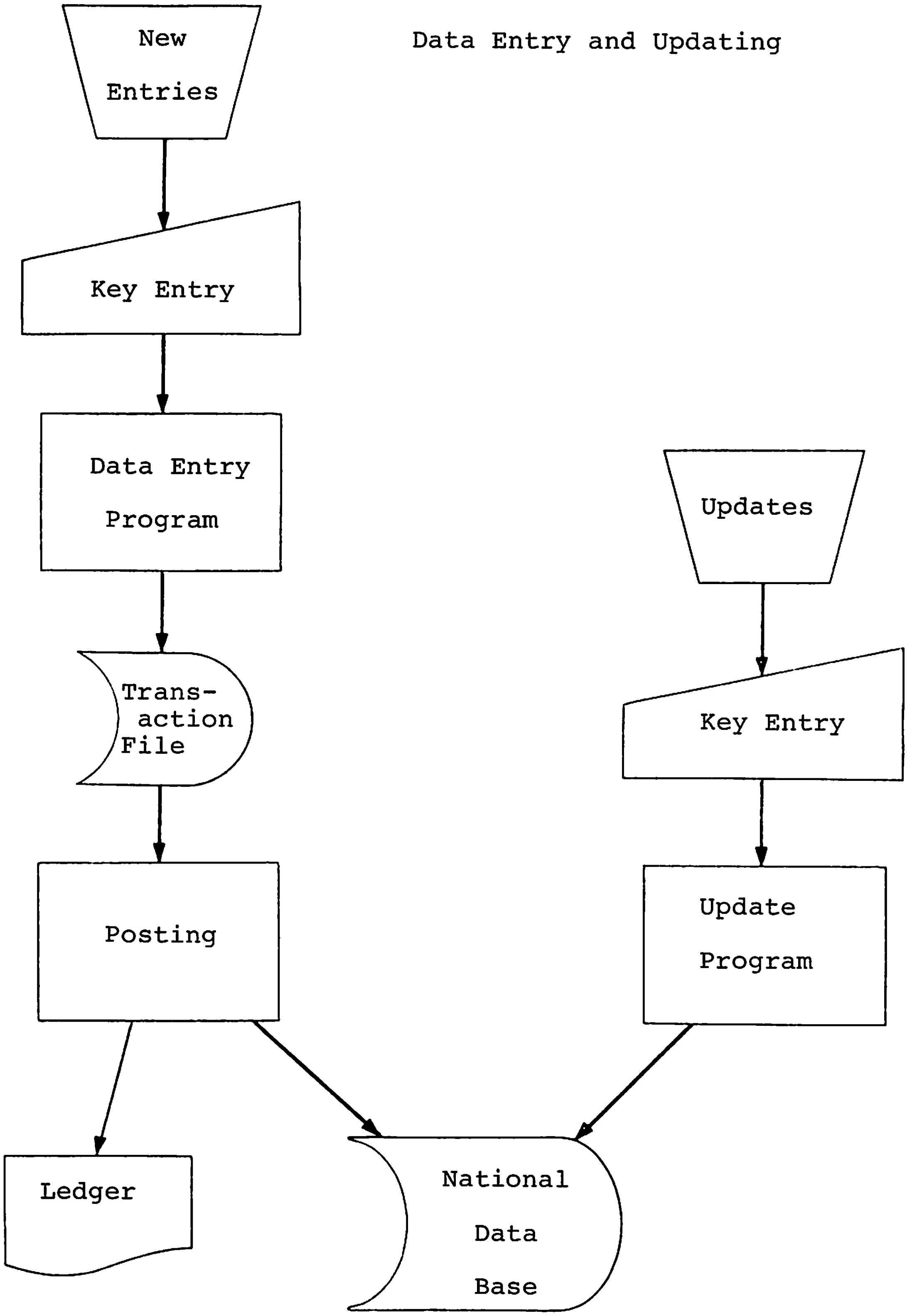

Data Entry and Updating

Figure 3.

Data Base Management Tools

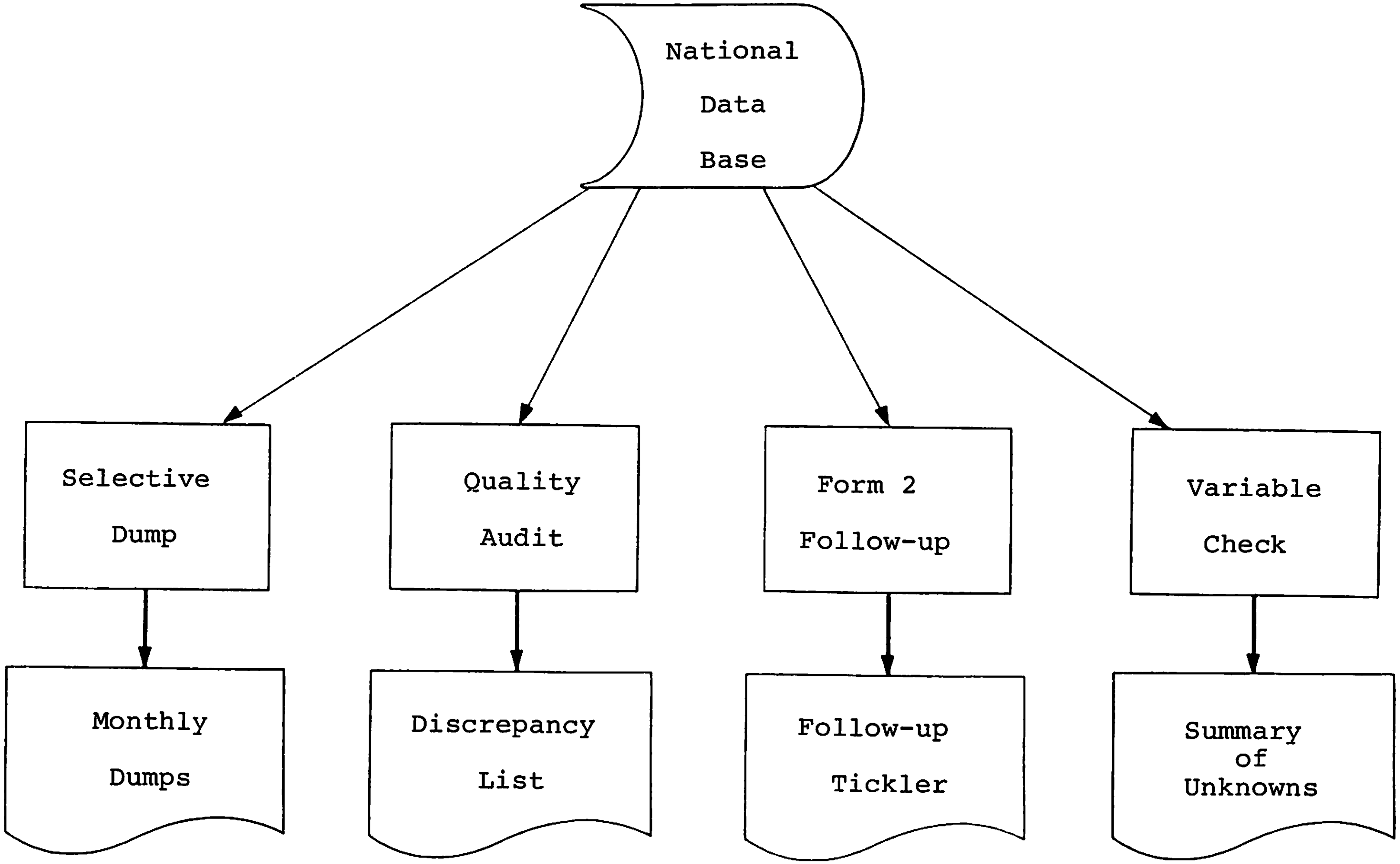


Figure 4.

analysis. As a result, there were no useful precedents for statistical inquiry and analysis.

With the growth in data volume, as well as the desire to perform statistical analyses and to make the data available for analysis via remote terminal, a hardware and software evaluation was conducted of available systems in the $100-150K price range. The HP3000 was selected as the best system in that price range. Selection was based on ease of use, multi-lingual capabilities, and data base management software in a time-shared and batch oriented system. NSCIDRC's data processing facilities are outlined in Appendix 2.

We elected to use IMAGE and, at least initially, QUERY. For our beginning efforts at inquiry into the data base and to check our conversion, QUERY was, to say the least, very handy to have. However, in establishing NSCIDRC's data bases under the HP-3000 IMAGE data management facility, it was obvious that the HP QUERY program, although a good general approach to on-line inquiry, was not adequate for our needs. As a result, NSCIDRC Computer Services undertook to design and implement a full function program that would serve all of our data bases and provide efficient interactive access to the data with our special needs for security, ease of use, and statistical analysis in mind. INQUIRY is the result of that effort.

In order to better comprehend NSCIDRC's data base management needs, let's examine the data base structure.

DATA BASE ORGANIZATION

NSCIDRC's data base structure was designed to anticipate the need for a variety of possible access modes. The present schema structure is outlined in Figures 5 & 6. Because the schema was designed to allow use of QUERY, certain inefficiencies (which may be obvious to the experienced HP3000 user) were introduced. We will address these aspects at a later point.

We utilized Automatic Masters because we anticipated access to the data base from a variety of directions: patient number, center, number of hospitalizations, anniversary year, etc. We have learned from experience that the only Masters we need are File-key (center, patient number and anniversary year, combined) and Center. The reason for this is that data is generally selected on the basis of a combination of logical selection criteria applied to several different variables or data items.

Data Base Structure

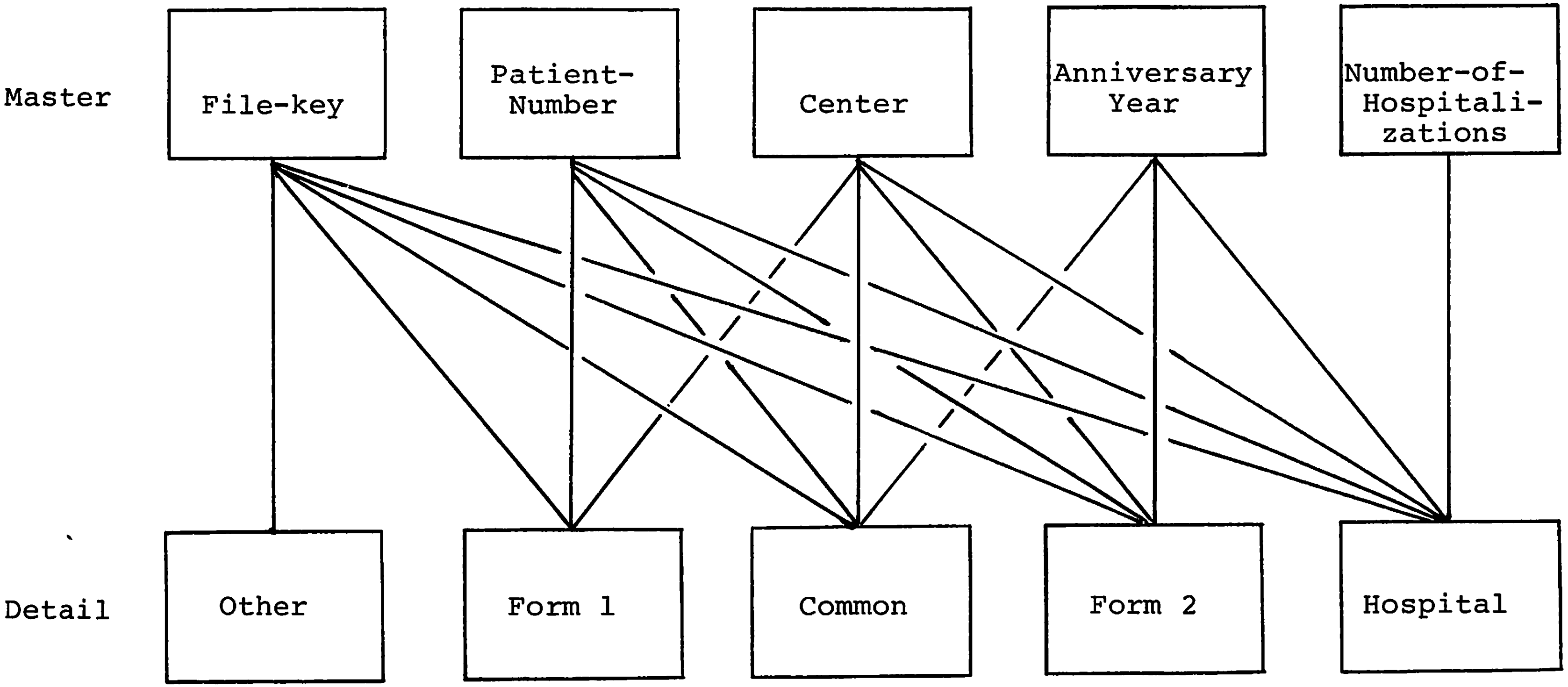
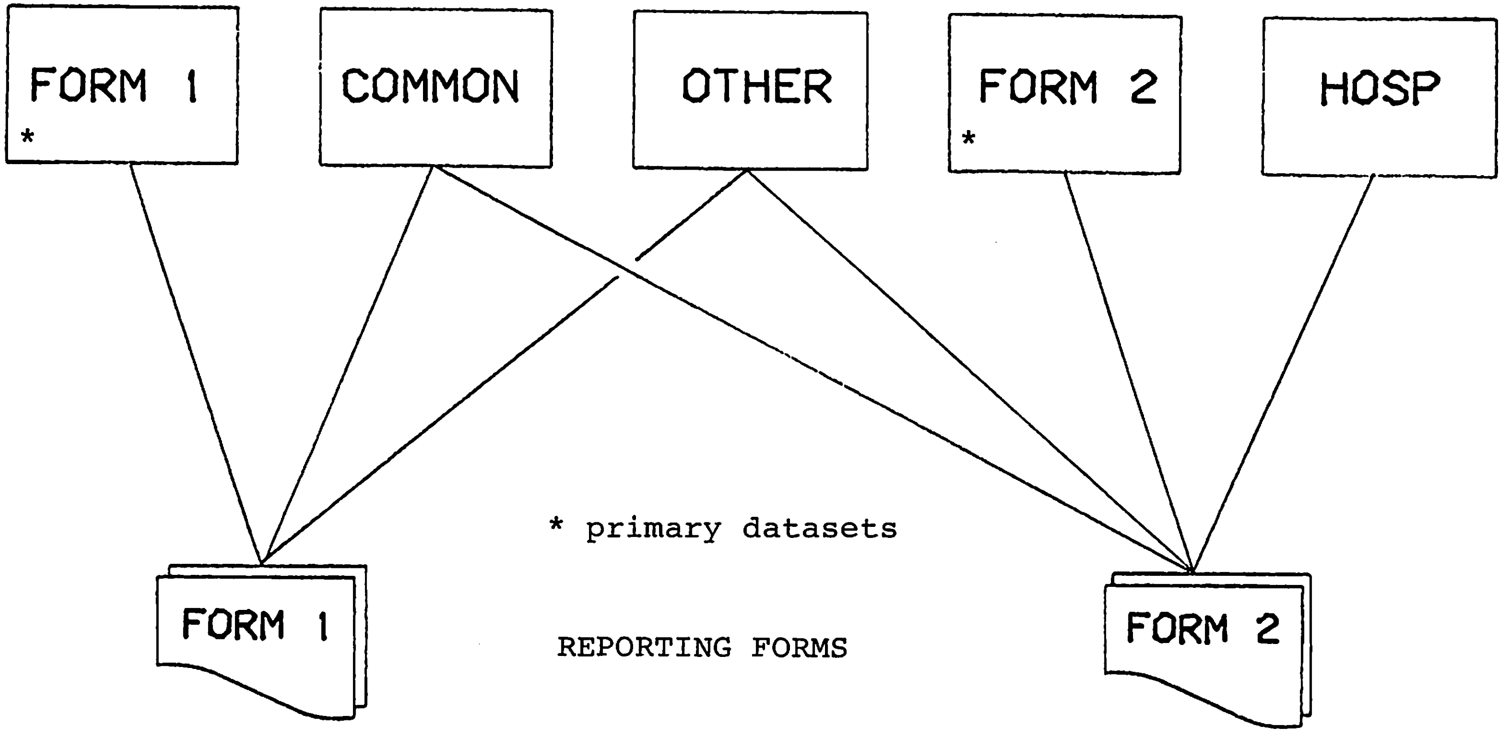


Figure 5.

THE NATIONAL DATA BASE



* primary datasets

REPORTING FORMS

Figure 6.

At NSCIDRC, the data bases are accessed in one of the following ways:

- * Serial, by primary dataset
- * Chained read, by Center
- * Calculated read, by File-key.

INQUIRY utilizes the three methods above, plus directed read.

RETRIEVAL REQUIREMENTS VERSUS QUERY

A brief look at the features of HP's QUERY program is appropriate here. These features are as follows:

- * Interactive English-language commands, like FIND and LIST
- * Single data set access
- * Command features such as:
  - Find command accesses any single data set for selection of data by logical comparisons
  - List command, with access similar to Find, produces columnar listings of desired variables; some column headings will be truncated
  - Report command allows flexible output report formats, including sorted details, column totals, and register manipulation (calculate averages, etc.)
  - Report commands may be repeated to display other variables (but List may not)
- * Update command features:
  - Direct access to a particular variable by name
  - Add, replace or delete a data record
  - Global replacement of a specific variable
- * Execute from a command file
- * Execute pre-written procedures for data selection, reports, etc.

While we liked QUERY's English-like command-driven style, we required multiple-dataset access. We wanted to be able to relate the entry forms and variables to the datasets so that the user need not be concerned with data base structure.

We wanted a List command that did some simple, automatic formatting, without losing vital header information. Since we were interested in statistical inquiry, we wanted elementary statistics on all listed numeric fields. Further, we wanted to be able to save a selected population and to extract from the data base as a separate file desired variables based on the subset population.

While QUERY must create an internal "tag" file of pointers to selected data, it is probably in an extra data segment. In any case, it cannot be saved, nor is it accessible to the user.

In addition, QUERY provides no simple means of creating an output disk file of selected variables from the subset population. The listing file can be equated, but that is a messy and undesirable workaround, particularly for the unsophisticated user.

DESIGN OF THE INQUIRY PROGRAM

The following outline summarizes the features of the INQUIRY program:

- Multiple data set access

- EDITable directory file contains indices to relate data sets, forms, and groups of variables; output formats for the List function; field type and position for the Find function

- Selection of variables by number as on the data entry forms

- Command features such as:
  - Find command accesses all data sets in the specified form
  - Find can pseudo-chain across form boundaries, that is, it can access both Form 2 and its corresponding Form 1.
  - Find can locate cases of multiple occurrences of a value (e.g. those patients with 2 spinal fusion operations)
  - Find allows use of parenthetic notation:

    ```
    F NATL1 V104 < 7 AND &
     (V120=" 030" OCC 2 OR V130=" 030" OCC 2)
    ```
  - Temporary Tag files created by the Find command define the population and may be saved and later recalled as required; both positive and optional negative tagfiles can be created
  - List command, with access similar to Find, produces neat columnar listings of desired variables; for numeric variables, produces elementary statistics at the end of the list; details may optionally be sorted, or suppressed; variables may be decoded thru automatic tabular look-up for more readable listings
  - List and Output File commands may be repeated to display other variables from the same population, and subsets from the population may

be Listed or Output to a File via the IF option which allows further selection from the "tagged" population
- Output File may contain up to 20 variables and is compatible with the input requirements of SPSS and LISA, statistical packages available on the system
- Frequency command produces a table of frequencies, cumulative frequencies, and cell counts, along with statistical totals.
- Transform function allows creation of a pseudo-variable for use in List, Frequency, and Output File commands.

* Execute from a command file

* Command termination on Control-Y

The relating of datasets, forms and variables was solved by the use of a driver directory file. All access to the data bases uses the directory file indices which are core-resident, except for the variable descriptors' portion which, because of its size, is accessed by a binary-search routine.

The use of a private directory file as a driver has many important implications. The data base may utilize single-byte fields, odd-length fields, multiple-occurring fields. Fields may be redefined. Thus a date may be accessed in its entirety as YYMMDD, or just the year as YY, while occupying only 6 bytes of space. The only limitation on redefinition is that search-keys must be uniquely defined, although even they may be redefined for purposes of data manipulation within the program. The directory file is not a privileged file and is quite separate from the data base. Therefore, it may be edited and modified as required without necessarily affecting the data base or its schema.

Because INQUIRY is designed for use by relatively unsophisticated users who are familiar with the forms and the patient data - but not data bases or programs - it assumes a set of operational defaults. These defaults may be over-ridden by a simple command. Among the defaulted options are: choice of single- or double-spaced listings, "noisy" or "quiet" mode (in which various messages are suppressed for the experienced user), and a lookup feature in which coded data is decoded for more readable listings (e.g. sex code of 2 becomes "Female").

All commands are accepted in both full English as well as shorthand, e.g. "FIND" or simply "F". Simple error messages and help messages are provided. Syntax is entirely free-form, with continuation lines and multiple commands on a single line allowed.

A CLOSER LOOK AT THE INQUIRY PROGRAM

INQUIRY is written in modular fashion in COBOL. While not strictly structured, it is functionally top-down in design. Although a large program, it has been carefully designed to minimize swapping and maximize execution efficiency. Its stack size of 3000 is necessary primarily because of an integral sort statement. Program segments responsible for the data base I/O are self-contained so that no segment transfer takes place until it has completed its task. As an example, the Find command is set up in one segment and executed in a second segment which retains complete control until that command is finished.

Data base I/O transfers are performed in "all-items" (i.e. full record) mode. The extraction of bytes is performed entirely by the program rather than by IMAGE intrinsics. This is the key to the odd-length field access, and the ability to handle multiple occurrences.

USING INQUIRY IN THE ANALYTICAL PROCESS

In practice, INQUIRY can be used to peruse the data base or to extract a data matrix to test a tentative hypothesis. The user will usually save his tag file which contains binary record pointers to the population selected. These pointers are used to perform directed reads in IMAGE. If the user should decide to extract a second group of data items, the tag file previously created provides rapid access to the same population. Figure 7 outlines the basic inquiry and analysis flow.

INQUIRY can be used interactively, although it is often more appropriate to create a Stream file for batch execution and return later for the results. This is because the time between responses to user commands may range from a few seconds to several minutes, depending upon the access mode, number of datasets and records accessed, and the overall system load at the time.

A sample Job stream with annotations is shown in Figure 8. A sample listing and frequency table are shown in Figures 9 and 10.

Our user base is spread over the United States and thus connect time and telephone charges can be expensive for some users. We have established alternate methods for the user with local analytical capabilities. The user can run INQUIRY and create a data matrix of selected variables from a tagged population. He may then elect to use SPSS or LISA on the NSCIDRC HP3000.

Inquiry and Analysis

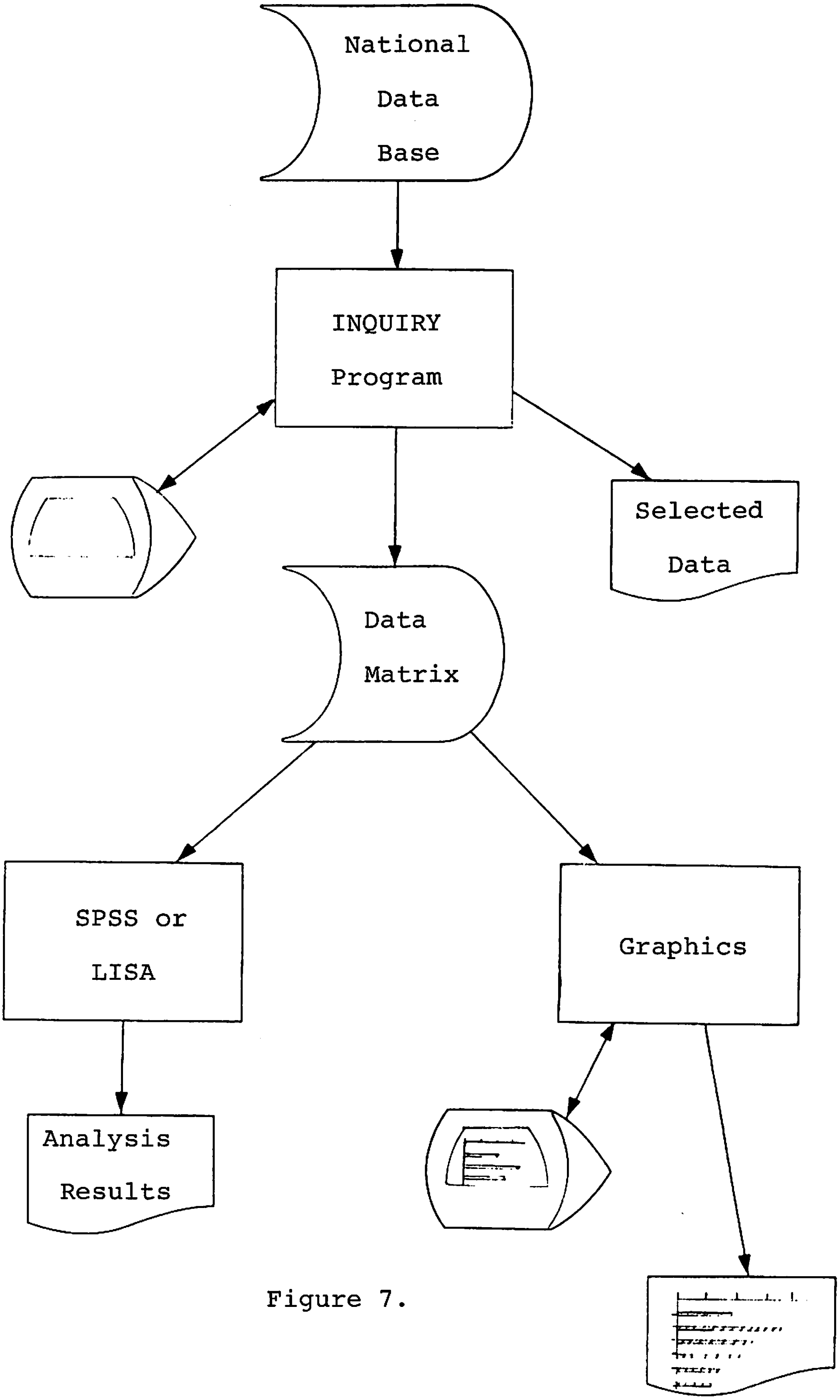


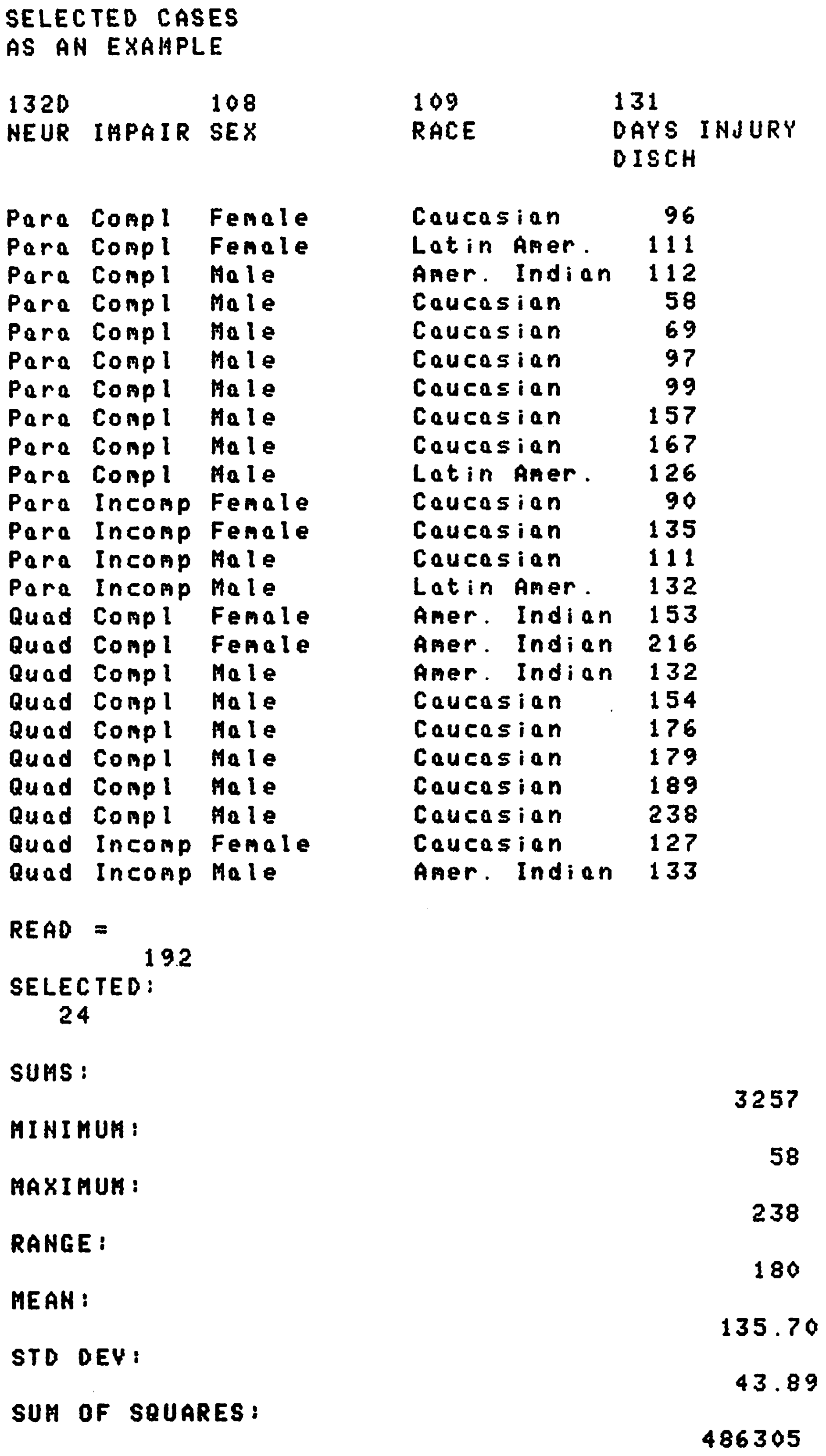
Figure 7.

```
!JOB MARTY.JEWEL/PASS
!RUN INQUIRY.MARTY.JEWEL
B NATL;M A T;M QUI;OL
                  specify which data base; access
                  all data; create a tag file;
                  use "quiet" mode; output to the
                  line printer
F NATL1 V103=1 AND V132D<5 (V119=7070 OR V129=7070)
                  in Form 1, select patients meeting
                  certain criteria
FR V107 15;S SORETAG
                  create a frequency table of variable
                  V107 (age) in 15-year groups;
                  save the tag file as SORETAG
T COST77
                  recall a previously saved tag file
TRAN V161+V166T
                  establish an equation for a
                  pseudo-variable (transformation)
LT V161 V166T TRAN
                  List totals only for the variables
                  shown, including the pseudo-
                  variable, TRAN
TRAN V160+V164+V169T+V170T
LT V160 V164 V169T V170T TRAN V172
OF CSTMATRX V161 V162 V163 V166T V167T &
 V168T TRAN V172 V132D V103
                  create an Output File called
                  CSTMATRX containing 10 variables
                  as columns, each patient being
                  represented as a row; from the
                  population given by tag file
                  COST77; "&" means continuation
E
!EOJ
```

Figure 8.

If the user has a terminal with a local storage capability (i.e. tape cartridge, diskette, etc.), the user can simply use FCOPY to transfer the data matrix to his terminal storage medium. Then the user can dial his local computing facility and feed in the data for statistical analysis.

For the user whose time constraints are more flexible, or who requires a large quantity of data, the data matrix (or the entire Regional Center data) can be dumped to a magnetic tape and mailed to the Center.



```
SELECTED CASES
AS AN EXAMPLE

132D         108          109          131
NEUR IMPAIR SEX          RACE         DAYS INJURY
                                       DISCH

Para Compl  Female       Caucasian       96
Para Compl  Female       Latin Amer.    111
Para Compl  Male         Amer. Indian   112
Para Compl  Male         Caucasian       58
Para Compl  Male         Caucasian       69
Para Compl  Male         Caucasian       97
Para Compl  Male         Caucasian       99
Para Compl  Male         Caucasian      157
Para Compl  Male         Caucasian      167
Para Compl  Male         Latin Amer.    126
Para Incomp Female       Caucasian       90
Para Incomp Female       Caucasian      135
Para Incomp Male         Caucasian      111
Para Incomp Male         Latin Amer.    132
Quad Compl  Female       Amer. Indian   153
Quad Compl  Female       Amer. Indian   216
Quad Compl  Male         Amer. Indian   132
Quad Compl  Male         Caucasian      154
Quad Compl  Male         Caucasian      176
Quad Compl  Male         Caucasian      179
Quad Compl  Male         Caucasian      189
Quad Compl  Male         Caucasian      238
Quad Incomp Female       Caucasian      127
Quad Incomp Male         Amer. Indian   133

READ =
        192
SELECTED:
   24

SUMS:
                                                  3257
MINIMUM:
                                                    58
MAXIMUM:
                                                   238
RANGE:
                                                   180
MEAN:
                                                135.70
STD DEV:
                                                 43.89
SUM OF SQUARES:
                                                486305
```

Figure 9.

```
AGE AT INJURY
IN 15-YEAR GROUPS

107
AGE


COUNT =
   31

 CELL VALUE    FREQUENCY   CUM. FREQ  CELL COUNT

   0  TO    14        16.13        16.13           5
  15  TO    29        41.94        58.07          13
  30  TO    44        12.90        70.97           4
  45  TO    59        12.90        83.87           4
  60  TO    74        12.90        96.77           4
  75  TO    89         3.23       100.00           1

SUMS:
        973
MINIMUM:
          5
MAXIMUM:
         75
RANGE:
         70
MEAN:
       31.38
STD DEV:
       19.57
SUM OF SQUARES:
    42031.00


MODE:
         15 TO        29
MEDIAN:
         26
```

Figure 10.

THE FUTURE OF THE DATA BASE

In the near future, we plan to reorganize the data base in order to eliminate the undesirable space-wasting aspects mentioned earlier. We estimate that the reorganization, while losing some compatibility with QUERY, will save approximately 13% of the disk space currently used. In addition, eliminating the Common dataset will reduce the number of accesses by 25 to 50 percent, depending upon the variables accessed.

The reorganization will take into account:

* Odd-length fields
* Multiple-occurring fields
* Redefinition of fields, including search keys
* Elimination of the Common dataset by expansion of the Form 1 and Form 2 datasets

Just how do we plan to implement the reorganization? A conversion program will write consolidated records to a tape, eliminating the Common dataset and the unnecessary bytes in various fields. Then we will purge the old data base, create the new root file and data base allocation. We will then sort the tape by file-key and utilize another special program to load the data base from the sorted tape.

Since we are also modifying the data base syllabus (definitions) and adding some new variables, the conversion program will have to translate some data to new values. This will result in a somewhat more complex program, but will allow us to do the job in a single pass.

The structure of the planned data base is shown in Figures 11 and 12 (compare with Figures 5 and 6).

SUMMARY

Spinal cord injured patient data from eleven Regional SCI Centers is submitted to the National Spinal Cord Injury Data Research Center in Phoenix, Arizona. The data is entered into an IMAGE data base on NSCIDRC's Hewlett-Packard 3000 using custom-designed software. The use of custom software for inquiry and retrieval was necessary in order to access multiple datasets at one time. It was also needed to relate entry forms and variables to the data base structure for user ease and convenience.

The INQUIRY program provides elementary statistics as well as the ability to save a subset population for later recall. It may also be used to create an data matrix as an output file for further analysis.

An added advantage of the INQUIRY program is the space savings which result from freedom from the usual limitations of the HP QUERY/3000 program.

Planned Data Base Structure

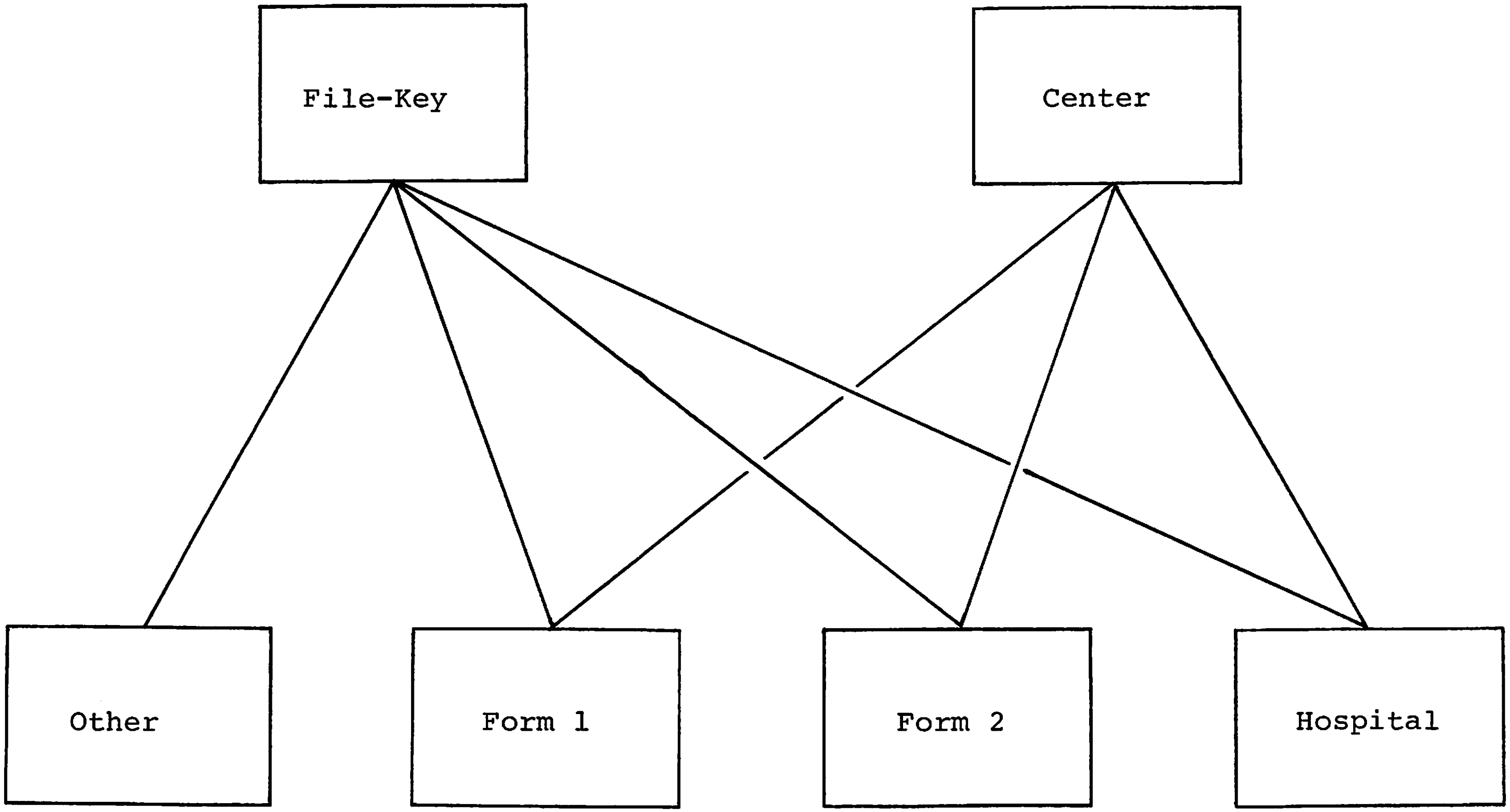


Figure 11.

THE NATIONAL DATA BASE

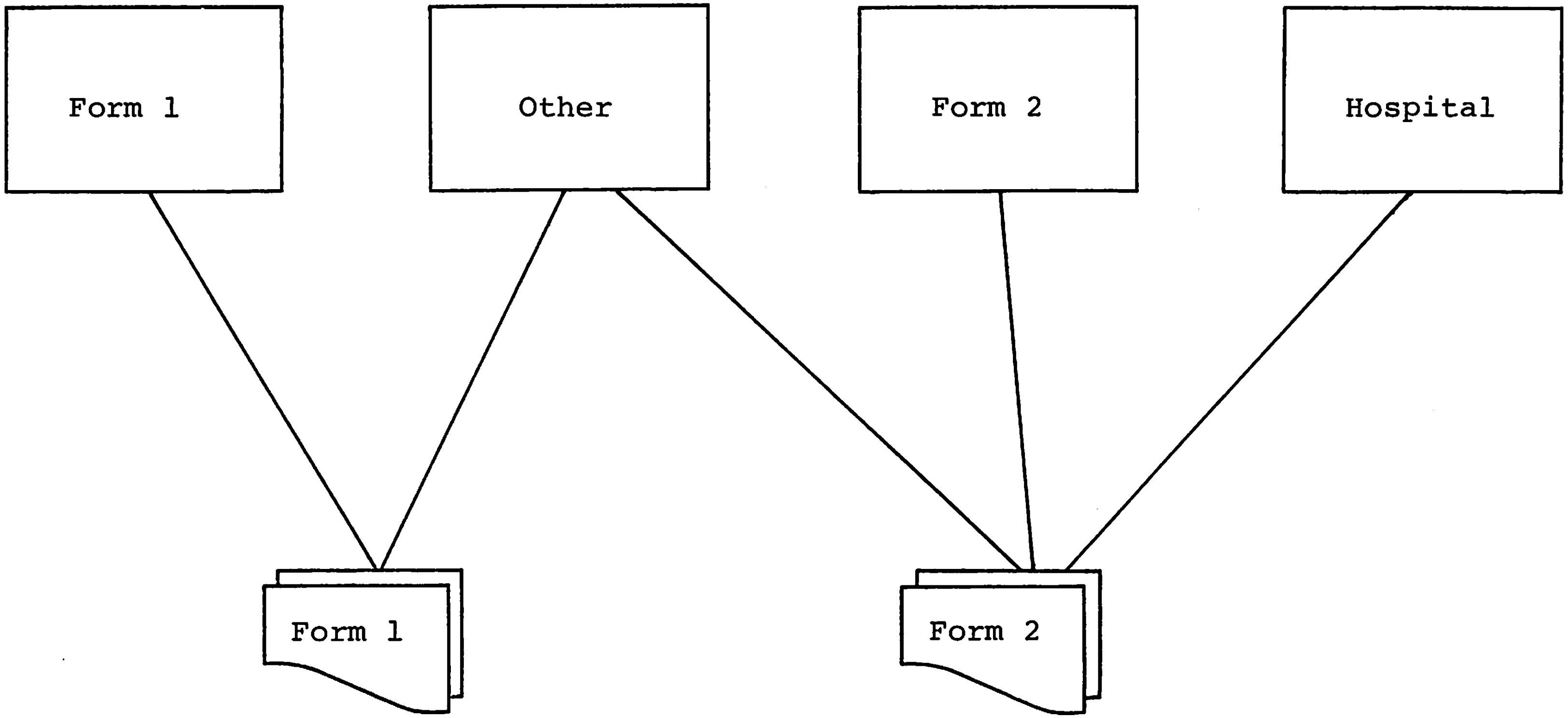


Figure 12.

The National Spinal Cord Injury Model Systems' Project is sponsored in part by the Rehabilitation Services Administration, Department of health, Education and Welfare. The following are participating institutions:

University of Alabama; Birmingham, AL

Good Samaritan Hospital - St. Joseph's Hospital; Phoenix, AZ

Santa Clara Valley Medical Center; San Jose, CA

Craig Hospital; Denver, CO

Northwestern Memorial Hospital - Rehabilitation Institute of Chicage; Chicago, IL

Boston University Medical Center; Boston, MA

University of Minnesota Hospital; Minneapolis, MN

Institute of Rehabilitation Medicine, New York University; New York, NY

Texas Institute for Rehabilitation and Research, Baylor University; Houston, TX

Woodrow Wilson Rehabilitation Center; Fishersville, VA

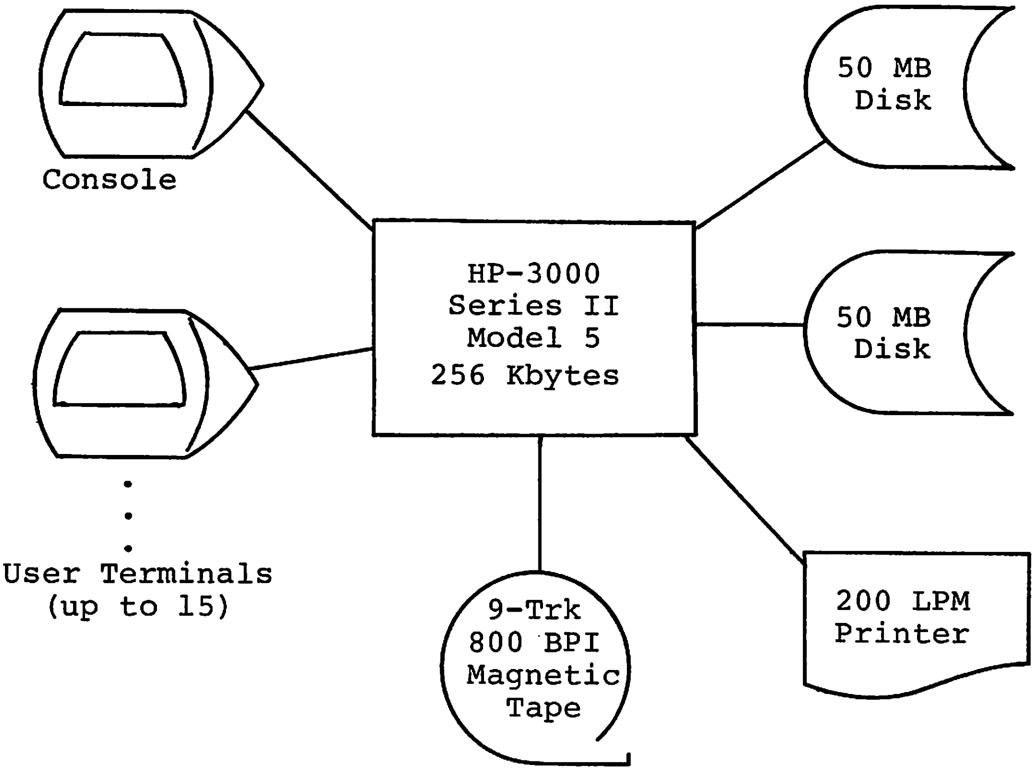
University of Virginia; Charlottesville, VA

University of Washington; Seattle, WA

NSCIDRC

The National Spinal Cord Injury Data Center

Data Processing Facilities



Software Support:

Languages: COBOL, FORTRAN, BASIC, SPL (extended ALGOL)

Data Base Facilities: IMAGE (data management system)
QUERY (inquiry and reporting)

Libraries: DEL (Data Entry Library)
Scientific Library

Operating System: MPE III (Multi-Processing
Executive III)