Downloading Problems?

by Kay Worrell[1]
Director, Survey Research Center
The Conference Board, New York

In the past two years, with the trend toward decentralization of many computer services and with the spread of microcomputers, new applications and new problems have arisen in research organizations. "Downloading," or moving files from a mainframe computer to microcomputer, offers some new opportunities and some solutions to old problems. But as with other solutions, new problems are also presented.

There are four areas in which problems may occur:

- The mainframe "source" of the data
- The data itself — its form
- The communications package and modem — the vehicle(s) with which the data are to be moved
- The receiver — the software (or system) into which the data are to be stored.

[1]Presented at IASSIST Meeting 1985, Amsterdam

In discussing some of the problems which might occur in these four areas, I will use two specific downloading applications attempted to solve Conference Board problems. One of these involved the downloading of numeric data from our own mainframe computer, a Burroughs; the other was a test downloading of mixed numeric and text data from an external on-line database. In both cases we were downloading into an IBM PC-XT.

As background, I should say something about the Conference Board and our research. The Conference Board is a not-for-profit business and economics research group, based in New York and with offices in Brussels. The Conference Board of Canada is headquartered in Ottawa. We are supported primarily by subscription income from Associate members and by conference fees.

My department, the Survey Research Center, processes questionnaires for researchers surveying business practices and economic trends, and assists research staff with computer systems and software. It is important to note that for most of our survey data, the "case" is the corporation or company, and the "background variables," or demographics, are company traits such as sales, assets, number of employees and type of industry.

Our specific applications

The first and simplest application was downloading coded numeric data from a dataset we had developed in our mainframe computer. The data was set up in 80-column records to be used with SPSS, as is most of the data we process from our questionnaires. The purpose was to provide research staff an opportunity to work on the data interactively in a PC statistical package, and to experiment with PC spreadsheet

and database management packages. The latter would offer more flexibility in using nonnumeric data such as names of companies, titles of indivuduals, and responses to open-ended questions. After downloading, we would add this information to the coded numeric files in the micro.

The second application was to try to download company information from an external online database as a possible source of background information for our questionnaire data. In the past this information was requested on the questionnaire, and checked upon receipt, or was added to the questionnaire after it was returned from printed sources such as the Standard & Poor's Directory of Corporations or Fortune magazine's annual List of 500 Largest U.S. Corporations. The information was then key-entered with the rest of the questionnaire data. Sometimes the information was encoded on the questionnaire label, prior to mailing, from our own mainframe computer list, where much of this information is also kept and updated regularly. Again, it would be re-keyentered into the specific dataset when the questionnaires were returned.

We decided to try to download company information from the Disclosure II database, available online through DIALOG. We wanted to download company names, sales, number of employees, and primary SIC (Standard Industrial Classification) number, used to indicate industry group. We were then going to explore ways to link this information with that in our data files, or organize it in such a way that it could be easily referenced by clerks in an off-line mode. The information in the Disclosure II database is derived from the forms that publicly-held companies in the U.S. are required to file with the Securities and Exchange Commission. Disclosure has an exclusive contract with the SEC to computerize this data, so that it is the most complete, authoritative, and up-to-date source available. I should note that this downloading was purely exploratory, and that permission would be arranged before the Conference Board would implement use of such an external source.

Thus our second application might offer solutions to several problems, by allowing us update our own mainframe list with the most current and reliable information, match this information with questionnaire data, and avoid considerable clerical work and redundant keyentry.

Interface considerations in downloading

Beginning in the mainframe dataset, the size of the file should be one of your first considerations. This will be influenced by the number of records as well as by the volume or size of each record. The critical question at this point is: Will the file fit on a floppy disk?

Format of the dataset is another important concern. Is it fixed field? SDF (standard delimited format) convertible? Is the data in column format? If so, is it SDF or DIF convertible? Is it numeric, alpha, or mixed? Are there multiple "lines" or "records (cards)" per case? Are there variable length fields or variable length records (a different number of fields possible on each)?

What are the host system characteristics? Are there line numbers? Is numbering an option? How long is each line, or record? What type, of end-of-line or end-of-record character is used? Can the host system transmit anything besides ASCII files?

What is the configuration of the communications hardward? Full or half duplex, synchronous or asynchronous line; speed of communication — baud rate? What communications package will be used? What will it do for you? Can you

move system files or just ASCII files? Will the receiving unit be a hard disk or diskette? How much space is available? What is the method of transfer, or "protocol?"

Finally, after the data is downloaded to the PC or microcomputer, there are the following questions. Can you load the information directly into the package you wish to use it with? Will it be desirable to load it into a word processing package? As an intermediary, for reformatting? For word processing uses? If it is to be used in a database management package or a spreadsheet package, will you wish to add additional data? To merge with other files?

The downloading

We used LinkIT to communicate between the two systems, and most often used KeepIT, a database management package, to receive the files. Both are produced by ITSoftware of Princeton, NJ and distributed by Martin Marietta Data Systems. The main advantage of KeepIT is its interfacing capabilities, allowing it to receive and reformat data and generate output directly in SDF, DIF or other ASCII format.

The steps to download are the same for all applications,and are:

- Sign into the communication package.
- Call up the host computer in which the "source" dataset is stored. (Set or reset communications parameters, if necessary. Most often these can be saved in the communications package used.)
- Sign into host computer and call up dataset. DO NOT ISSUE A "LIST" OR "DISPLAY" COMMAND YET. You may want to check on the size of your dataset before downloading.
- Indicate to communications package that you wish to "receive" a file. this will probably be done with a function key. (LinkIT leads the user to such an option with a menu.)
- The package will ask you to name the file. This may be done using normal naming conventions, including designating the disk drive address. Be sure there is sufficient space available for the file.
- Return to host system and issue a "LIST" or "DISPLAY" command. The "listing" should then be "received" by the communications package. Note: If the host system allows an option of listing "unnumbered," that is, without line numbers, use this option. Otherwise you will want to remove the line numbers after the file is downloaded.
- When the listing is completed, the downloading will be completed. Most often you will see this indicated by the end of a count of characters received appearing on the screen.
- Sign off the host system. Then exit the communications package.
- The downloaded file will be stored on your hard disk or diskette, with the name you supplied in step 5.
- You may then load the file into the package of your choice. It may be called directly into most word procesing packages. If you use a database management package such as KeepIT or dBase III, you must first define the file parameters, including all fields and the length and type of each.

The first problem we encountered downloading data files from our mainframe was that of the line numbers, mentioned above. This can be avoided by using an "unnumbered" option. Another, related problem was some unnecessary information repeated on each "card" of a case or observation — questionnaire identification number and "card" number. this information took up the first nine columns on each "card" or line, and was easily removed using a word processing package. Or, loading the whole file into a database management package, that information could be defined as "dummy" variables or fields, to be omitted later.

We have tried to keep the size of data files downloaded rather small — less than three "cards" per case and only a few hundred cases. Larger files are not very efficiently processed in a PC. Even so, downloading can be time-consuming at normal baud rates of 300 or 1200. To speed up the communicating of larger files, our EDP department provided us with a special line to transmit data to the PC at 9600 baud. This required changing the modem from Hayes to "direct" and some changing of the plugs, besides changing the baud rate setting.

Also, as noted above, it is important to estimate the size of the file to be received and to be sure that the disk on which it will be received has sufficient space. If this is not the case, you may lose much of the data you tried to download, and waste considerable time.

As our data was in 80-column records and was already in fixed-field format, we expected no problems defining the fields to the database management packages. Using KeepIT's menu, we just "READ DATA." We discovered, however, that KeepIT requires each record to have 80 columns, and no less. Some of our records were shorter than 80 columns, and CANDE, the Burroughs operating system command-and-edit language did not fill those records in with anything recognizable by KeepIT. So our staff had to place a character in the 80th column of each record before downloading into KeepIT. This was done with a "replace" command in CANDE, but it meant we could only use 79 columns for data.

We have used KeepIT as an intermediary, to format files for dBase, Lotus 1-2-3 and WordStar's MailMerge. In dBase, after defining the file parameters, we can use "APPEND". If the file contains addresses, we can use either KeepIT or dBase to create a MailMerge input file. We can produce a DIF output file with KeepIT, using SPREADSHEET INTERFACE, and then "Import" this file into Lotus 1-2-3. Or we can use the dBase-Lotus interface for this purpose.

To add data to any of thse files, we need only create the additional fields, copy existing data in, then edit the file to add the new data. Or we might create parallel files with a linking ID number for the new data, then merge the files.

On moving data into statistical software packages, we found some common good points and bad points. In their favor is the fact that virtually all PC statistical packages seem capable of accepting ASCII data files. We have experimented with SPSS-PC, StatPac and StatIT. Some time can be saved if SPSS is to be used on the PC, as much of the labelling used for SPSS in the mainframe can be downloaded and adapted for use in SPSS-PC. Of course if you are using an IBM mainframe with Kermit, downloading will be even more convenient.

A disadvantage is that PC statistical packages have a smaller capacity, as would be expected, than mainframe packages. Thus only partial files could be used. Another problem encountered was that StatPac will not read multiple-line or multiple-card files. The data must be in a continuous stream ending with a carriage return after each record. A record may not exceed 255 characters. A utility file, which comes with StatPac, must be used to concatenate 80-coumn records to form a StatPac-readable

record.

Trying to download data from an outside system created considerably more problems. As was noted at the beginning, this was an experimental application. Again using LinkIT, we signed into DIALOG, and then into Disclosure II. We discovered there were three basic forms in which data could be presented from Disclosure. The first did not include the variables we wished to see — sales, assets, number of employees and SIC number. The second, the Corporate Resume, contained considerably more information than we wished — about two full screens of data for each company. The third contained much, much more — whole company records, including text from annual reports. We decided to try to download the Corporate Resume for a small number of companies, just to see how long it would take. Selecting only companies with upwards of $40 billion in net sales, we narrowed the search to 17 companies. We then downloaded these companies. Even at 1200 baud, it took nearly 10 minutes to complete this download.

Based on the time it took, and especially considering the line charges for DIALOG and Disclosure, we determined that this would not be an efficient updating mechanism. Further inquiries revealed that DIALOG will, under certain circumstances, download large files for the user onto a 9-track tape, to user specifications. We also found that we could purchase the entire Disclosure database on 9-track tape from Disclosure, including updates. (And since this paper was presented, I have received promotional information on MicroScan, a software product from DISCLOSURE to assist the PC user in searching and downloading from that database.)

Other problems that became evident, but which we did not seek to solve after the downloading, were the very large size and variable length of each record.

A positive factor we discovered about Disclosure was that, in addition to virtually all the company data that is publicly available for corporations, each record contained the D & B identification number, as well as that used by Standard & Poors' and others. Ticker symbols used on the stock exchanges were also included. Thus data from this database could easily be merged with data from any other, where one of these common numbers were in use.

On the basis of this experience, we have decided to consider purchase of Disclosure on tape to update our mainframe files. We have also decided to use one or more of the common identification numbers listed in Disclosure as an identifier on all future datasets we plan.

I should add that, although I will not discuss it at length here, we have also found it convenient to enter data into the PC on occasion, and upload it to the mainframe. We have done this using Lotus, to allow us to compute new variables during the data entry step. It should be noted that for large datasets the reaction time for Lotus becomes quite slow. We then uploaded one "card" of data and merged it with nine other "cards" that had been keypunched in the traditional way. We have also entered some questionnaire data directly into a "screen" set up in dBase III, and uploaded some of this information for use with SPSS. The problem we had to deal with in this type of transfer was transmitting a pause after each record, or "line," to allow the system time to assign a line number for each.

It is expected, of course, that future developments in specific software packages will include interface enhancements. These, along with improvements in communications hardware and software, will greatly facilitate our ability to move data among packages and systems.■

LinkIT -- screen one

```
LinkIT 1.2    (c) Copyright 1983 VM Personal Computing    OFFLINE
                   (c) Copyright 1983 IT Software

              Your PC  ID is: THE CONFERENCE BOARD INC.

              F1 = Call a Computer Named CONFBD

              F2 = Answer a Call from A PC

              F3 = Review the Directory of Computers

              F4 = Set Personal Computer Options

              F6 = Edit a File

              F7 = Edit a File

              F8 = Run a Program

              F9 = Stop Printing

              Esc = Exit       F10 = HELP
```

LinkIT -- screen two

Directory of Computers

Name	Telephone Number	Speed	Type	Notes and Comments
A PC		300	PC	IBM PC using LinkIT
COMPSERV		300	Host	COMPUSERV service
CONF BD	83,456	1200	Host	THE IN-HOUSE BURROUGHS SYS
DOWJONES		300	Host	
SOURCE		300	Host	SOURCE timesharing service
TSO		300	Host	Direct call to a TSO system
TYMSHARE		300	Host	TYMSHARE or equivalent
UNATTEND		300	PC	Leaves LinkIT Unattended
VM		300	Host	Direct call to a VM system

Use PgDn and PgUp to scroll the Directory

F1 = Call Name at Cursor F2 = Answer Name at Cursor

F3 = Add a New Name in Directory By Copying Entry at Cursor F4 = Review Connect Options for Name at Cursor

Esc = Quit F10 = HELP

LinkIT -- screen three

```
LinkIT 1.2      Your PC ID is: The Conference Board      ONLINE

              F1 = Return to Terminal Screen

         Alt F1 = Redial or Reanswer the Telephone

         Alt F2 = Hang up and Return to Main Offline Menu

              F3 = Send Files to Another Computer

              F4 = Receive Files to Your PC

              F5 = Set Current Connect Options

              F6 = Edit a File

              F7 = Print Files
              F8 = Run a Program

              F9 = Stop Printer or File Transfer

                     F10 = HELP
```

KeepIT -- screen one

```
RECORDS: 0            KeepIT - MAIN MENU            FILE: C:DAY


        DATA ENTRY                    FILE DEFINITION

  ED - ENTER DATA              DF - DEFINE THE FILE
  PD - POST DATA               DC - DEFINE CONSTRAINTS
  RD - READ DATA               DI - DEFINE INDEXES

        DATA MAINTENANCE              INTERFACES

  CD - CHANGE DATA             CI - CalcIT INTERFACE
  VD - VIEW DATA               GI - GRAPHICS INTERFACE
  CF - COMPUTE/FILL DATA       SI - SPREAD SHEET INTERFACE
                               ML - MAIL LIST INTERFACE
        REPORTS                FI - FORMS INTERFACE
                               ST - STATISTICS INTERFACE
  PR - PRINT A REPORT          WO - WRITE OUTPUT FILE
  TR - TABULATION REPORT       SH - ShowIT INTERFACE
  SR - SUMMARY REPORT
  WI - MOVE REPORTS TO WriIT

        QUIT                          HOUSEKEEPING

  QF - QUIT FILE               DO - DIRECT PRINTER OUTPUT
  QD - QUIT TO DOS             FM - FILE MANAGEMENT
  QA - QUIT TO ASKIT           MC - MAINTAIN CATALOGUES

          PLEASE SELECT AN OPTION:
```

KeepIT -- screen two

```
FIELDS: 0              DEFINE THE FILE                  FILE: C:DAY

                  A - CREATE FIELDS FOR A NEW FILE
                  B - INSERT A FIELD
                  C - DESIGN/EDIT FIELDS ON SCREEN
                  D - DELETE A FIELD
                  E - DISPLAY A FIELD
                  F - SCAN THE FIELDS AND MAKE CHANGES
                  G - DISPLAY THE DATA ENTRY SCREEN
                  H - PRINT THE FIELD SPECIFICATIONS
                  I - PRINT THE DATA ENTRY SCREEN
                  J - SET FILE PARAMETERS
                  M - RETURN TO MAIN MENU

                  PLEASE SELECT AN OPTION:
```

KeepIT -- screen three

```
FIELDS: 1          DEFINE THE FILE          FILE: C:DAY

                SPECIFICATIONS FOR FIELD:  1


   (1)   PROMPT   1-                (6)    TYPE OF FIELD
   (2)   NAME                       (7)    MAX LENGTHT
   (3)   page,pageend                        (8)    LOWER LIMIT
   (4)   ROW                        (9)    UPPER LIMIT
   (5)   COLUMN                     (10)   INPUT SPEC

                                    (11)   FORMAT
   (12)     DEFAULT/FORMULA
```

KeepIT -- screen four

```
RECORDS: 0          MAIL LIST INTERFACE          FILE: C:DAY

                A - WritIT/MULTIMATE
                B - WORDSTAF/MAILMERGE
                C - EASYWRITER/EASYFILER
                D - PEACHTEXT
                E - WORDPLUS-PC
                F - WORDPERFECT
                G - EDIS+WORDIX
                H - SPELLBINDER/EAGLEWRITER
                I - QUOTE MARKS/COMMA DELIMITED
                J - FIXED LENGTH

                PLEASE SELECT AN OPTION:
```