Utilizing Mainframe Data on PC Platforms: Problems, Solutions, and Techniques

by Carol Wickenkamp[1]
WAE

Introduction
As more organizations and institutions downsize computer facilities in order to make greater use of the ubiquitous and inexpensive desktop computer, the problem of how to get non-ASCII data from there to here becomes increasingly common and pressing.

Archival requirements as well as data utilization are affected by the platform shift; additionally, users are expecting greater access to data than in past decades and devising access methods with and without the blessing of the MIS staff. Life expectancy of archival tape media from the 70's and 80's is diminishing. All of these issues draw us to ask the question: How do you get the data off the mainframe and onto the computer.

Let us break the big problem, the great need, into smaller and more manageable problems, in the spirit of the eating of the elephant.

Problem: Determining whether the data is even suitable for conversion
Careful evaluation of the data will help you determine whether to shelve the project or to move onward. This information is critical in determining not only feasibility, but potential cost of the project. This evaluation will help you to discover those unpleasant exceptions to the rule that will require expensive programming and special processing that can drive costs for conversion out of the feasibility range.

Techniques:
1)Check the physical condition of the media itself, especially if has been many years since cleaning and copying of the tape

2)Try to evaluate the adequacy of documentation, so far as record format, field definitions and descriptions, and code tables.

3)Do your best to determine that this data is really what you thought it would be, that it is suitable for your needs, or that isn't already duplicated elsewhere in a more accessible format.

4)Determine the tape density on older tapes, and for very old tapes, whether they are 7 track or 9 track[2].

5)Non-EBCDIC data formats crop up on older tapes especially, and can greatly increase the effort and expense of conversion. Look for packed decimal, zoned decimal, packed bit, or binary data, these data formats will need special conversion techniques[3].

6) Unusual file formats will also need special conversion techniques: for example, tapes from military sources may be in NIPS, those from medical facilities may be in MUMPS, and PICK systems have been in wide use for many years[4].

Solutions:
If your facility has mainframe to PC connections, your tapes are in good condition, and your tapes are readable by your current mainframe or mini facility, you can run a sample of 3 to 5 megabytes from each file across the network. The PC interface cards necessary for the PC to mainframe connection will automatically convert mainframe EBCDIC data to ASCII data. This sample data will help you to determine the adequacy of the available documentation and the presence of unusual data and file formats, which we will discuss further in this section. Data which does not convert directly from EBCDIC to ASCII can be readily identified. Even in very large files, a sample of this size will almost always yield usable data in all fields.

If, however, your data is truly historic, Just accomplishing this task can be a problem in itself. Unless your MIS staff is familiar with older computers, tape, and data formats, this evaluation may be better left to professionals. The section on Data Conversion Service Bureaus addresses the issue of older tapes.

Data Conversion Service Bureaus

There are data conversion service bureaus in most cities that deal with old data on a regular basis. For very old and fragile tapes, consider contacting a disaster recovery service; many of these agencies have the techniques and equipment to do serious data recovery. Be prepared to pay for this initial evaluation, and ask for a quote (based on the number of files you'll want evaluated). You will

need an evaluation that will cover all the points discussed above, in 1 through 6. In addition to the initial evaluation, request a quote for providing a 3 to 5 megabytes sample EBCDIC to ASCII conversion from each file if the tapes are readable. Unless the files are under 20 to 30 mb, ask if they can take two small samplings (500K), one from the middle of the file and one from the end of the file as part of the 3 to 5 mb sample, and find out how much extra it will cost you for these small samples. Current price for EBCDIC to ASCII conversion is usually about $10 per megabyte. Get cost quotes for your evaluations from more than one agency, and also ask if you can contact previous customers, as you would for any contract service.

Tape Drive Peripherals

If you have neither mainframe to PC capabilities nor the funding for service bureau work, or for other reasons have decided to tackle the project in-house, consider rental or purchase of a 9 track tape drive that will interface with a PC. These tape drives will come with software that will perform simple EBCDIC to ASCII conversions, and some will have software will have software with even more capabilities. For example, Qualstor's drives come with software that will convert directly from EBCDIC to Dbase. Drives are available that will handle varying tape densities; Overland makes a tape drive that will handle even the very old 800 bpi density as well as the contemporary 6250 tapes. If you know that your tapes are not fragile and you can safely run them, you can use a tape drive peripheral to do your initial evaluation of your data, running the same 3 to 5 mb sample. Data conversion service bureaus often rent drives, as do some of the larger computer equipment rental companies. The cost is usually about one teeth the purchase price; drives adequate for most conversion jobs will rent for around $600 per month.

Documentation and Identifying Unusual Formats

Using the documentation you've gathered, and a print out of your sampleASCII data (start with just a few records), you can begin the task of reading the raw data. This process will uncover gaps in your documentation as well as "funny" data. Frequently unusual data and File formats will be easily discovered on initial examination, before you even begin to check your data against the documentation. See Figure I for "Funny Data"; the fields that contain the curly brackets signal the presence of zoned decimal numeric fields, as do "/" characters and unexpected periods. Zoned decimal will be converted incorrectly in a simple EBCDIC to ASCII conversion, as is obvious.

Other non-EBCDIC numeric formats can also yield exotic results.

If you find no indication of problem data, use the field descriptions in your documentation to mark off the Fields in your data, as in Figure 3. Check your data fields one by one against both the field definitions and code table, if some of the data is coded. Here in Figure 3 we have clean data, with names, dates and Julian dates, cities, etc. where they should be and in the proper format.

Make sure that the code values in coded fields are represented in the code tables. Should you find codes that are not listed in the code table, but the rest of the data is clean and in agreement, you have probably encountered either an undocumented code (if there are many occurrences) or data entry errors. If you have undocumented codes, you can sometimes extrapolate the meaning from the data when the entire file is converted. Often a further search for more documentation is necessary. (Both the National Archives and NTIS retain copies of some Federal computer documentation.) Lack of sufficient documentation can doom your conversion project, unless you can be satisfied with either converting the portions of the data that you can identify, or just archiving the data in the hope that you can obtain the requisite documentation at a later date.

Take samples of 20 to 50 records from different places in your 3 to 5 mb sample and verify the data. If your are able to obtain records from the middle and end of your life, be sure to check them, as sometimes another file with a different format was appended to the first data file. Should you find evidence of multiple files, you will want to make a note of it so that when you have the tape converted, the data can be run off in separate files during the conversion process.

Determining Conversion Costs

Using the evaluation information about your files, you can begin to calculate costs. For example, if you send 300 mb of clean EBCDIC data to a data conversion service, and they charge $10 per mb, your charges will be $3000. To this you must add the cost of target media sufficient to store that volume of data. This figure will of course vary according to the media. Should your facility plan to download the data from a mainframe to a PC, your in-house costs will, at a minimum, include target media costs and computer time, which may or may not include computer operator charges. Coordination with your MIS department will be essential in defining costs for in-house conversion. If you have data that requires special processing, costs may include data recovery fees for very old and fragile tapes, or programming costs to convert data that is in non-standard data or file formats. You will need to obtain a second round of quotes for this

work, which will be more expensive than standard conversions, or negotiate with your MIS department for programmers to do the work. Doing the conversion yourself, for those without mainframe connections or funding for service bureau work, will be addressed in section Converting Data on a Low Budget.

If your data will require the programming services, expect to pay a minimum of $50 per hour. Programming costs in major metropolitan areas will be greater. As with other contract work, obtain more than one quote and ask to speak with previous customers. Try to speak with customers whose programming and conversion needs were similar to yours, in order to ascertain that the programmers have actually dealt with this type of data or file format; you don't want to pay for the programmer's learning curve.

Problem: Converting data on a low budget
There are those facilities who will not have the resources of an eager to help MIS department, or the budget to cover thousands of dollars for data conversion services. There are alternatives that can put the data conversion and migration process in the realm of the possible for even the most underfunded facility.

Techniques:
Hardware
Before we begin the "hands on" process of converting this data, we must have some repository for the finished product. Depending on the volume of data, there are a number of target media that will be appropriate.

High capacity hard drives are becoming very affordable, with prices dropping to around $1 per mb and even less for very high capacity drives of over 1 gigabyte. This drop in price put desktop mass storage within the reach of low budget facilities.

The lowest cost storage media will be the inexpensive PC backup tape. QIC 80 tape, which is becoming a standard for entry level backup, will store 250 mb of compressed data; this means that you will usually be able to store more than 250 mb of data on one tape. The drives are inexpensive, currently selling for under $200, and will function very well in older AT class PCs. The media will cost about $20 per cartridge. The drives are adequate for short term archival storage (not recommended for a permanent solution), but are slow and inefficient if you plan to use the data frequently.

Removable media hard disk drives are available in either internal models or portable models that interface with the PC through the parallel (printer) port; these drives offer another attractive alternative. Prices on these drives rapidly dropping; at the present, a drive in the 110-120 mb range can be purchased (with some judicious shopping) for about $400, including one cartridge; higher capacity drives are available. Each cartridge contains a hard disk platter, and the user can easily switch cartridges. The media costs are about $65, and prices should fall rapidly. The advantages include very fast access to data for those who need frequent access and portability. These drives can be compressed with disk compression utilities, increasing the storage potential. They are an excellent choice if your data files are in the appropriate size range and you will require frequent access to the data.

DAT backup drives are more expensive starting at about $1000, but they are very fast, they store gigabytes of data, and the cartridges cost about half as much as the QIC80 cartridges.

Solutions:
Data Copy by Data Conversion Service Bureau
Service bureaus will make an exact copy of your data and write it to your media. The current cost for this service will be in the range of $1 to $1.50 per mb of data. For example, if you are using QIC80 tape, request the bureau to make a copy of the data file(s) onto QIC80 cartridge media, which you will then restore to a hard drive at your facility for do-it-yourself data conversion, or simply retain as archival storage. (A discussion of do-it-yourself data conversion will follow in this section.)

If you are using a tape backup medium, be sure to tell the service bureau the name brand of your tape drive, as cartridges written by one brand of tape backup equipment be readable by equipment manufactured by another company. It is wise to do a test run with a trial tape cartridge written by their equipment, to determine whether your equipment will read the tape. You will also want to request that the data file tape headers (preliminary system information written when the tape file was created) be stripped from the data, and that only data be copied onto your medium. If you have a large number of tapes, It will be wise to pre-determine a meaningful data file naming scheme, so that you will know which data file is which when you get them back.

Nine Track Tape Drive Rental

Your facility may decide that tape drive rental is the most feasible course. Basics on PC peripheral 9 track drives were covered in an earlier topic. The company that rents you the tape drive may provide both installation and removal of the interface card if you have no one on site who can do it. As was earlier discussed, the software that comes with these drives will provide the option of converting the EBCDIC data to ASCII as it is copied off the tape and onto your storage medium. Those who are

not familiar with tape conventions such as blocking, and fixed and variable length records, determine the degree of customer support available from the rental agency You may need some initial instruction. If you have no special conversion needs, this is a most cost effective solution to the data conversion.

Data Conversion Software

Service bureaus that do data conversion and rent 9 track tape drives often sell special data conversion software that has more features than the software that is bundled with their tape drives. Typically, software of this type will handle the unusual data formats mentioned earlier, and can convert standard variable length records to fixed length records. Expect to pay $200 and up for this software. Do not count on conversion software to accomplish the task of converting the non-standard file formats discussed earlier; you probably will still require programming services.

Frequently the software interface is intimidating and may be hard to get used to, but the conversion process itself is not overwhelming. Generally, you will be required to mark off the data fields (as you did with your sample, only on screen rather than on paper) and then define the conversion process that is to take place, i.e.. EBCDIC to ASCII, binary to ASCII, or packed decimal to ASCII. When you have defined your conversion instructions, your file is ready to be converted by the software.

It is a good idea to run a partial conversion of 500 to 1,000 records to verify the accuracy of your field definitions. Sometimes the process will require several tries before all the bugs are out of your conversion instructions, and it is far faster to convert 1,000 records for a sample than to convert 100,000 records. The speed of conversion will depend upon the processor speed of your computer, the complexity of your conversion instructions, and the length of your records. You can use the measure of 1 megabyte per minute as a rough rule of thumb. Although most of these programs will operate on files residing either on the tape drive or a hard disk, it is much faster to copy your file onto a hard disk and do the conversion from disk.

Problem: The data is so heavily coded that it will be difficult to work with

As a rule, database programming relies heavily on code table to hold frequently used values; old mainframe data can be coded in every field, thus yielding very compact files. The code values were replaced at processing time so that reports were understandable. This sort of data is very cumbersome to use, even with modern and easy to use database programs such as Paradox, Alpha Four, Access, etc.

Solution:

Given the low cost of hard disk storage, It is becoming more feasible to simply replace the coded fields in databases with their values, yielding a significantly larger, but easier to use flat file database. Even with a two or three fold increase in file size, this solution can bring comprehensible, easy to manipulate data to the most unsophisticated user. It is far faster and more accurate to extract reports or meaningful data screens from a database that contains “Lutheran” rather than “07”, “Buick” rather than “15” or “CA” rather than “05”.

Expert programming skills are not necessary to accomplish these replacements, a moderately skilled in-house programmer should be able to do the job. Even if it is necessary to hire a programmer, it should not be a major expense, unless you have a large number of heavily coded files.

Conclusion

Although moving data from older mainframe generated tapes to a PC platform is a process that requires planning and attention to detail, the task is not insurmountable, nor is it always exceedingly expensive. With the exception of very old or non-standard tapes, much of the work can be done in-house and with a small budget, utilizing moderate computer skills.

Notes:

2. 7 track is an obsolete tape standard which used the 6 bit BCD (Binary Coded Decimal) code together with a parity bit. The contemporary 9 track drives will not read 7 track tapes.

3. Although data conversion software renders these numeric formats harmless to the non-technical user, a discussion of these formats is included for those who are interested. Numeric data format which will not convert in a standard EBCDIC to ASCII conversion include:

Packed Decimal with low order sign bit

This is the normal IBM packed decimal field.

Zoned Decimal with low order sign bit

This format is generated by some COBOL, PL/I and Assembler systems; although not common, it is still in use in some contemporary installations. Zoned Decimal is a standard EBCDIC numeric character field with the exception of a sign code in the high order nibble of the low order byte, with C hex and F hex being a positive sign code and D hex a negative sign code. This results in invalid EBCDIC characters in the low order byte of some zoned decimal fields.

Binary with most significant byte first

This is the format in which IBM mainframes normally

process binary data; normal PC binary format is binary with least significant byte first

Packed with high order sign bit

This is a binary format with the sign bit in the high order nibble of the high order byte.

Packed with no sign bit

This is a normal packed field, except that all nibbles contain a significant digit (no sign field) and the field may begin and/or end on a nibble boundary.

4.These are all non-standard variable length file formats. MUMPS has been widely used in VA hospitals and in medical clinics, and is still common. PICK usage extends across the commercial spectrum. NIPS was designed specifically for use on IBM 360 computers, and is no longer in use.

Sources:

* For further information on tape formats, labeling and file conventions, you can contact:

American National Standards Institute, Inc.
1430 Broadway, New York, NY 10018.
Tel : (212)6424900

Ask for publication X3.27, “Magnetic Tape Labels and File Structures”

IBM tape labeling conventions are explained in the IBM publication “OS/VS Tape Labels” (GC26-3795-3, File No. S370-30) and “DOS/VSE Tape Labels” (GC33-5374-1).

DEC information is described in “Guide to VMS Files and Devices” (AA-LA06A-TE), available from DEC.

* If your facility is not in a metropolitan area, you may find several reputable data conversion service bureaus advertised in PC Magazine, which is available in most drug stores and supermarkets.

* Two companies that produce data conversion software, each with different capabilities, are:

NovaStor
30961 Aguora Road, Suite 109
Westlake Village, CA 91361
(818)707-9900
Fax (818)707-9902

Overland Data
5600 Kearny Mesa Road
San Diego, CA 92111
(619)571-5555
Fax (619)571-0982

Service bureaus may also have information on other data conversion software.

1. Paper presented at IASSIST 1994 in San Francisco.. Reprints of this paper are available from: Carol Wickenkamp, WAE, PO Box 349, Clarkston, WA 99403