DATA USE AND REUSE

The movement and use of survey data across different hardware and software platforms.

Peter Wills

Mercator Computer Systems Ltd, 5 Mead Court, Cooper Road, Thornbury, Bristol
BS35 3UW, United Kingdom

Paper presented to ASC International Conference – September 1992

Abstract

The movement of survey data between different hardware and software platforms is increasingly being requested by computer users. This paper reviews the techniques that are currently available, and proposes new methods of standardisation between different software packages.

1. INTRODUCTION

During the 1960's and 70's, survey data was held almost exclusively on mainframe computers. The choice of storage media for moving data was limited to punched cards and magnetic tapes. The choice of hardware was limited to only a few manufacturers, and the number of software packages could be counted on the fingers of one hand. The fact that data could not be moved across hardware and software platforms was generally regarded as only a hindrance, as most Data Processing houses specialised on a single mainframe computer, and their teams of "spec writers" were trained to use a single software package.

With such incompatibility existing in the industry, it is little wonder that punched cards thrived for so long – the information was visible, the cards could generally be read by other computer systems, and with the aid of a trolley the data was portable.

As hardware developed in the 1970's and minicomputers arrived, software houses altered their programs to operate on the new hardware, but the structure of the industry remained unchanged. Companies still had computer departments, with teams of computer operators, and the user departments "communicated" with the computer department via systems analysts. These bright young people (or failed programmers) then prepared indecipherable specifications for the computer programmers, hidden away behind closed doors. Users were always kept at arms reach, would virtually never operate the computer and would certainly never actually write computer programs for themselves.

The 1980's brought the microcomputer and the ubiquitous PC. The potential then existed of the user not only operating a computer, but even writing computer programs. In those early years, every microcomputer was incompatible with every other microcomputer, particularly when it came to data storage and data transfer. By 1986/87, the industry realised that the future lay in the PC, with IBM compatibility ensuring that PC computers became a standard tool in both office and home environments.

Those software houses with roots in the mainframe market have ported their software to PC's and generally retained compatibility across hardware platforms in terms of data structure and data movement. PC software houses have each used different data structures and where necessary have developed import/export routines to link to other software packages. The result is now a wide range of well-respected survey and statistical software packages, at prices to suit all budgets and performances to match.

The survey and statistical computing industry is now sufficiently mature that it should give greater consideration to user needs. Users are increasingly computer literate, and have need to use more than just one software package, and are quite capable of doing so. However, data structures still remain unique to individual software packages, with no common interchange format across software and hardware platforms, and arrive at a common format for passing data between systems. The result will benefit all – software houses will sell more packages, and users will make greater use of survey and statistical techniques. Whilst DOS on PC computers has led to a wide range of sure interfaces, the Windows environment provides a common interface and will undoubtedly be a topic that dominates the computer industry in the 1990's. A window of opportunity now exists to create a common interface for survey and statistical data, and it should not be missed.

2. PUBLISHED PAPERS ON THE MOVEMENT OF SURVEY DATA

Considerable work has been carried out on the generalised procedures for file transfer, with *Hale (1981)* carrying out a comparison and critique of various file transfer protocols. Other separate works by *Salzbery, Loomis and Folk* have studied the principle of file management structures. Much energy has been concentrated on the development of relational and hierarchical database structures, but no consideration is given in any of the publications to problems of moving data between different software platforms.

3. MOVEMENT OF DATA ACROSS HARDWARE PLATFORMS

Movement of data across different hardware platforms involves establishing communications protocols, and transferring exact copies of data from one computer to another. The most widely used method of data transfer between mainframe computers and other computers is a software package called *Kermit*. It is generally considered as public domain software and is consequently widely distributed by hardware and software suppliers.

During the early 1980's most new personal computers were based on the CP/M and DOS operating systems, but each manufacturer created its own floppy disk format. Consequently, most personal computers were incompatible with each other. A UK software product, *Swap,* overcame this by providing an implementation of the software for each different computer format, and supplied with a suitable cable, *SWAP* was capable, for example, of linking a BBC under its own operating system, to an IBM PC and PCDOS. A total of 150 floppy disk formats were available, consequently the permutations were virtually unlimited.

The dominance of the IBM PC compatible removed the need for products such as *SWAP*, and *Laplink*, with its increased transfer speeds, became the industry standard method of transferring data between different disk formats. Its widest application has been in the transference of data between computers with different size floppy disk drives, and linking desktop PCs to laptop PCs, with the latter having a hard disk drive, and perhaps no floppy disk drive at all. A PC/Apple Laplink version now enables users to transfer PC data to Apple computers and vice versa.

Digital Equipment Corporation, or DEC, developed its first PC computer in 1982, named the Rainbow, as both a link to their VAX and PDP range of mainframe and minicomputers, as well as being a PC computer. The Rainbow failed to become established as a free-standing PC computer, because of its pricing, poor compatibility and lack of distribution into the dealer networks, and after a number of false starts, the project was dropped. More recently, DEC have forged links with other computer manufacturers such as Apple to provide users with a closer interface between mainframe and PC computers, enabling easy transfer between hardware platforms.

Within the PC environment, the vast majority of computers are IBM compatible. The one notable exception is Apple, a manufacturer that has created niche markets and generated considerable loyalty amongst its users. Few software packages are available for both formats, because there are fundamental differences in programming techniques between Apple and PC's. The solution for many Apple users is to purchase an emulation package, such as *SoftPC* that will enable an Apple computer think that it is a

PC computer. Whilst there is some degradation in terms of speed, operation of the software on the Apple is identical to the PC. A similar version for the *Next* computer has recently been released.

4. A MOVEMENT OF CASE/RESPONDENT DATA BETWEEN SOFTWARE PACKAGES

Software houses generally devise methods of data storage, with consideration to their own particular hardware problems, and their own particular programming techniques. Such an approach is perfectly logical and ensures speed and efficiency in program operation. The result is that many mainframe packages use column binary formats and many PC packages use ASCII formats.

However, implementations of column binary formats are many in number, as are implementations of ASCII. The result is that no two packages are likely to be compatible in the storage of their case data.

Cynics might argue that software houses have a vested interest in keeping data specific to their own system, as it hinders users from migrating to other software packages. Horror stories do exist of software houses charging large sums to export data out of their software, only for the next software house to charge large sums to import the data into theirs.

Most reputable software houses and data processing bureaux have import and export links within their systems to read and write data to and from other software packages. However, the routines are generally not universal, and are specific to just a few other software packages.

The survey analysis system, *Merlin*, provides users with a powerful *Toolkit* option, to read and write data to and from a number of sources. It will handle ASCII data up to 10,000 characters in length, as well as column binary (80 column cards), in many formats including IBM format, Quantime format and binary/ASCII (where each byte contains 6 "punches" treated as a number).

An alternative approach has been taken by *Caloxy*, a software house in California, who have seen a business opportunity in the fact that each software package maintains its own file structure. It has developed a software package called *DBMS/Copy*, that supports over 60 products ranging from databases, (such as Dbase, Paradox, Dataease), spreadsheets (such as 1-2-3, Excel, Quattro), Statistical packages (such as SPSS, Minitab, Systat, BDMP, Glim, SAS) and Time Series (such as Soritec, Autobox, Forecast).

DBMS/Copy works on the principle that it has knowledge of the data format and file structures of its 60 host systems. It then imports data in one format and exports it in another format. It does this by referencing a table specification for the file to be imported, converting it in memory into a standard DBMS/Copy format, and then be referencing a table specification for the file to be exported, writes out data in the new format.

The approach taken is very logical and well documented, and rules are carefully explained for those situations when it is unable to convert particular strings of data. The software will even accommodate data split across a number of 80-column cards. An extra module, *DBMS/Copy Plus* has been added to allow users to have greater control in the conversion process. For example, users are able to select particular subsets for transfer, and to carry out arithmetical operations as part of the transfer.

It is salient to note that of the 200 software packages listed in the *Software for Statistical and Social Survey Analysis 1992-93* directory compiled by the SGCSA, only 22 packages are included in the of 60 packages supported by DBMS/Copy. All are US based, which is not surprising since *Caloxy* is itself based in the US. It is also apparent that DBMS/Copy has to be able to handle different versions of the same software product. If the file structures change between different versions, a separate implementation is required for DBMS/Copy. Consequently, software such as DBase appears under Dbase2, DBase 3 and Dbase 4. However, the product does have a place in the market, and one is surprised to find that only 300 sites existed in 1991, according to the SGCSA Directory. Distribution in the UK has now been taken over by SPSS, who have produced a comprehensive user manual, and the exposure of this product is likely to be increased in the near future.

It should be stressed that whilst DBMS/Copy performs an admirable function in respect of the conversion of case/respondent data from one format to another, the data dictionary aspect is just as important, but is handled with less success by the software.

5. MOVEMENT OF DATA DICTIONARY DEFINITIONS BETWEEN SOFTWARE PACKAGES

DBMS/Copy is capable of converting the data dictionary, but is limited to the variable name, variable type and format, and missing values. It will also accommodate variations between packages in terms of different variable name lengths and naming conventions. It will calculate the maximum size of a variable name in the output format, and truncate where necessary, readjusting to ensure that no duplicate variable names exist. It will accommodate variations in the use of special characters such as @ . and – and replace any inconsistent values with an underscore(_). DBMS/Copy will not, however, translate variable codes and labels between different packages. Many of the database and spreadsheet packages do not themselves accommodate codes and labels, whereas it is essential within survey analysis software. For example, a two digit variable called *Age* would be translated correctly to any other format, but codes of 0-18 with a label of "Youth", 18-25 with a label of "18 to 25" etc would not be translated. This could be considered a major limitation of DBMS/Copy in respect of survey analysis software requirements.

Direct links between survey analysis software packages have been developed with considerable success. Mainframe users of *Research Machine/Star* from *Pulse Train Technology* had made requests to be able to analyse the same survey on PCs under *SNAP* from *Mercator*. The data dictionary and respondent data already resided on the mainframe, and the plan was to transfer both datasets to the PC. A common interchange format was agreed between Pulse Train Technology and Mercator, with Star exporting its data to the common format and SNAP subsequently reading it. The consequence was that neither package needed to amend its file structures, and in the event of upgrades being made to either Star or SNAP, no amendments by the other software house were required, so long as they could still handle the same common interchange format. To ensure integrity in the transfer of the data dictionary across different platforms, a check digit was incorporated at the end of the file.

This approach could be handled by other combinations of software houses. Alternatively, they could agree a common interchange format, and formulate rules for handling variations between individual software packages.

6. PROPOSALS FOR NEW STANDARDS IN DATA TRANSFER

To generate standards for the movement of survey and statistical data across different hardware and software platforms, there are three alternatives:

1. All software packages could adopt an identical file structure. This would be impossible to achieve and would undoubtedly stifle any progressive software development to the point of working at the level of the lowest common denominator.

2. Each software package could develop a direct link to each of the other software packages. At first glance, this would appear straightforward, but firstly, many software houses would be unwilling to divulge too much information to potential competitors. Secondly, in the event of a software upgrade to either package, an amended link would have to be released. Supporting users in this aspect could be problematic.

3. Each software package could retain its own file structures, but export its respondent data and dictionary to a format common to other software packages. The format would have to be capable of handling both the largest and smallest problem size, as well as the least and most sophisticated functionality. In the same way as the software package DBMS/Copy has rules and regulations, this common format would need to be carefully planned and regulated.

   It is proposed that the major suppliers of survey analysis and statistical software be approached to ascertain their views and establish if there is sufficient will to cooperate. The major software suppliers in the UK are often the authors of the software, with the remaining companies supplying software from the USA. It would be logical to establish standards for UK and USA software first, before considering software from other sources.

   Areas of consideration would include:

   - Naming conventions for variable names
   - Variable types and formats
   - Naming conventions for respondent data files and dictionary files
   - A single standard format for writing respondent ASCII data
   - The requirement for the name of the source software package, version and release to be stored at the beginning of the data dictionary file

The benefits to users of such an approach is that they would no longer feel restrained by their software, knowing that they would easily be able to move data between packages for particular operations.

The benefit to software houses is that they would be likely to increase their customer base, and generally sell more copies of software. It is generally accepted that Apple computer users run more applications that PC users because of the easy transfer of information.

One obstacle preventing users from using a number of different software packages has, until now, been the software interface used. Each package has its own user interface, and each one can take some time to get used to. The Windows environment will change this and provide a common user interface. Consequently, the need for a common system of transferring data between different survey and statistical packages is not just timely, it is now urgent.

REFERENCES

1. R.W.S. HALE File Transfer Protocols – comparisons and critiques National Physics Laboratory (1981)

2. B. SALZBERY File Structures

3. M.E.S. LOOMIS Data Management File Structures

4. N.E. MILLER File Structures using Pascal

5. Study Group on Computers in Survey Analysis (SGCSA) Software for Statistical and Social SurveyAnalysis 1992-93 Compiled by C. ROWE, A. WESTLAKE and P. ROSE