The Promise of Multimedia: Data for Every Computer

by Janet Vavra[1]
Inter-university Consortium for Political and Social Research

When the microcomputer arrived on many of our desks in the 1980's, few of us realized just how dramatically this relatively small (often mysterious) piece of equipment would change our lives. Not only did the microcomputer alter the way we perform our daily tasks, it literally changed the way we view and interact with the world. Sending messages to colleagues half a world away was something we did either through Western Union or the postal system; data were shipped on magnetic tape through the mail. One looked for a response to an overnight letter in several days, assuming the marine life did not nibble through the undersea cable in the meantime. Today with the use of email facilities and public data networks, individuals send messages across the country and around the world with the same ease as they once picked up the phone and placed a call across town. Today many users have desktop computers (workstations) that have more computing power than many mainframes had 10-15 years ago, and the machines do not require an entire floor to accommodate them and their peripherals.

The user of a microcomputer, or workstation, can work with an enviable array of hardware and software performing data transfers, database searches, accounting tasks, data analyses, and receiving and sending messages without ever leaving the office. A researcher could conceivably conduct a project from beginning to end from the keyboard (or with a mouse) of a micro.

By searching the library's on-line catalog, relevant publications can be identified thereby eliminating long hours spent in the library by the researcher or a graduate student. Searches are not necessarily limited to local libraries, but rather identify publications available from a variety of sources. On-line databases can be used to locate machine-readable data containing the needed variables. If access to such databases is limited to selected users, a request asking for a given search could be sent via email to the individual authorized to search the databases.

Data could be ordered through electronic mail or electronic ordering systems. In some instances data could be downloaded through public data networks directly to the user's hard disk or to a local data library facility and then through a local area network to the user's machine.

Analysis can be performed with micro-based software and the results shared with project colleagues who may be at other locations, again using email or public data network capabilities. Small files of information can be exchanged through Bitnet, while larger files can be sent through networks such as Internet.

Finally research reports can be prepared with wordprocessors, many of which have graphics capabilities. Frequently the ease with which changes can be made in any document can create a problem of another sort: one cannot resist making another "final" change or one more addition to the document.

One could go on citing similar scenarios in almost every field and activity. The point is that many of the capabilities that were only available to users at central computing facilities, or not at all, are now available on the powerful machines many people have in their offices. Needless to say, this has impacted on organizations, such as the Inter-university Consortium for Political and Social Research (ICPSR), that provide data and related services to many of these users.

This paper will try to identify some of the changes brought about by the advent of the microcomputer. It will primarily look at those changes that impact both on researchers and on those organizations that provide machine-readable data to researchers and instructors. It will seek to identify some of the ways in which these organizations will and may provide their services in the future. The focus will be primarily on the challenges faced by the ICPSR both in continuing to serve users with more traditional computing environments and those at institutions where the activities of the central computing facility have been largely replaced by personal computers.

While the past two to four years have seen a growing interest in what can be termed "alternate media", data continues to be transmitted and exchanged among facilities on magnetic tape. However, the day is rapidly approaching where magnetic tape will not be the medium of choice but rather the medium of last resort. Magnetic tapes continue to be a medium that generally requires a mainframe. This is at a time when many central computing facilities are starting to cut back on services they have traditionally

provided, and many are also cutting back on staff that support these services. The reasoning frequently is that mainframes will serve as giant gatekeepers and servers while the microcomputers will take over the day-to-day computing needs of users.

With the central computing facilities cutting back on individual user services and users finding themselves with impressive computing power on their desks, it is natural that demand will increase for data and other supportive services that are compatible with micros. However, given the variety of configurations one can have at the micro level and the differences in individual preferences, it is no easy task to come up with products that will meet everyone's needs. Additionally archives face the very painful reality of the very high cost of converting all of their holdings from a mainframe-compatible format to a primarily microcomputer-compatible one.

Almost without exception, microcomputers have floppy diskette capabilities. However, not only do the diskettes generally come in two different sizes, they also can each be written in different densities. (Sounds a bit like the old days of seven- and nine-track magnetic tapes.) One can try to identify the format used by the most users and write diskettes routinely with those specifications. While that may be the only thing that makes sense for an organization that supplies data to thousands of users each year, there will remain those users who absolutely cannot use the standard product and must have data at different specifications. It seems to make sense to have a standard product that most users can work with and to deal with those users that cannot handle the standard product on a case-by-case basis.

Unfortunately, while floppy diskettes are an excellent medium for transmitting small collections of data, problems begin to arise when large data collections that contain more than a couple of megabytes of data are involved. One solution is to supply the data in compressed format. Most of the compression software around reduces the size of a file from 70%-80% of its original size. Since the capacity of most diskettes ranges from an average of 350 kilobytes for a 5 1/4" low density diskette to 1.4 megabytes for a 3 1/2" high density diskette, it is easy to see that large files, even when compressed, very quickly become impractical for this medium. Supplying one data file on numerous diskettes can create problems for software that may eventually have to manipulate the data. Large compressed files have to be decompressed and space has to be available locally to accommodate the decompressed data.

While the user has to be concerned with the amount of space available on their their microcomputer in order to be able to work with the data arriving on diskette, the data producer is further concerned with the effort that must go into preparing the data for diskette. It may be necessary to reformat the data to make it compatible with the micro environment. For example, PC-based software frequently cannot accommodate large record sizes with ease. Additional problems may arise with large numbers of cases and/or large numbers of variables. Depending on the work that needs to be done, reformating could be an expensive proposition. Accordingly, it may be necessary to identify only certain collections that can be provided on diskette and further to routinely provide these data in only a selected number of diskette formats.

Optical media go a long way toward solving storage problems when it comes to large collections of data. One of the more popular optical media is the CD-ROM. On a disk no larger than a 5 1/4" floppy, a CD-ROM can easily hold over 600 megabytes of raw, ASCII data. While access on a CD-ROM is slower than with a floppy, many producers bundle the data on their CD-ROM with software which helps to reduce the retrieval time. In other instances, users are not concerned about the length of the retrieval time, since time spent on their microcomputer does not result in direct costs the way time spent on a mainframe does. Instead they set up a batch job to run on their micro during the lunch hour or overnight.

Despite the high capacity of a CD-ROM and some of its other attractive features, the CD-ROM is basically not an inexpensive medium. Users usually must purchase a CD-ROM drive. Generally a CD-ROM's performance depends both upon the drive and driver used and on the power of the machine on which the work is being performed. It may even mean purchasing a different micro, if the current micro is not suitable for CD-ROM applications.

From the producer's point of view, a CD-ROM product can be a very expensive undertaking. If the data are to be

bundled with software, the producer must either identify existing software and then seek licensing agreements to use the software, or must write in-house software. Licensing agreements can be costly; the preparation of in-house software may, however, involve an even greater financial commitment. Data will additionally need to be prepared for input into the software or may need to be restructured for the microcomputing environment. Another alternative is to not supply any software with the product and leave it up to the user to identify software to be used with the data. This latter approach is more akin to using the CD-ROM as a data transmittal and storage medium than as a complete data transmittal, storage, and retrieval system.

While data can also be compressed on CD-ROM, the large volume of information that can be stored on CD-ROM usually necessitates special retrieval software for full or partial extracts. It is easy to visualize the problems that could arise if a user had to decompress a 550 megabyte file stored on CD-ROM onto a hard disk, or other local storage media, before being able to manipulate the data.

After the decision has been made regarding the nature of the CD-ROM product, premastering and mastering must be done before copies can be made. Normally premastering and mastering is done by service bureaus although producers can opt to purchase the necessary equipment and software for in-house capabilities. The charges for such capabilities preclude most organizations from deciding to master their own CD-ROM products. Generally the costs for producing a CD-ROM are such that only selected data collections can be considered for the medium.

The transmittal of data over public data networks has a great deal of appeal to both the users and the data producer. By simply identifying the data needed and giving the appropriate set of commands, the user can theoretically transfer any data collection needed in a matter of minutes. This can all be done without any direct intervention by the producer; the producer need only be notified in some electronic manner that the transaction occurred. As network speeds have been increasing from T1 (maximum 1.544 megabits or 200 kilobytes per second) to T3 (maximum 45 megabits or 590 kilobytes per second), this mode of data exchange has created a great deal of interest. But as with all of the alternate media discussed so far, there is good and bad with this option.

It is very attractive from a user standpoint to be able to simply give a few commands on your desktop and have megabytes of data arrive over the lines in a matter of minutes. There is no need to wait days for an order to be processed and then to always be concerned that it will not arrive in time either for a paper deadline or class assignment. It would eliminate the waiting that takes place when the user discovers that another data collection would have been a better choice than the one originally requested.

It certainly is true that if public data networks worked in practice as they sound in theory, our data exchange problems would be over. However, some of the same problems that impact other media are also at work here. The speed with which data arrive over the lines is the result of a number of factors, including the different routes they must pass through to get from the source to the user's machine. The speed with which the data make that journey will be only as fast as the slowest link along the network path. Therefore, users never actually experience the maximum data transmittal speeds quoted for any given network. While every effort possible is made by the public data networks to assure complete transfer of data sent, transmittal problems can arise, resulting in incomplete transfers. The machine receiving the data must have space to accommodate the information coming down the lines. Finally, it is likely that not all data formats will readily lend themselves to public data networks. For example, since most of the data going over the lines are ASCII, EBCDIC binary data such as that found in OSIRIS dictionary files and other similar formats will certainly not be usable on the receiving end.

After looking at each of the several media available and the advantages and disadvantages of each, one may very well ask which is the best approach. We at the ICPSR have been spending a fair amount of time exploring each of the different formats. Unfortunately, we have not found any simple solution that will provide everything for everyone. Instead, we have concluded that we will be fortunate if we can provide something for everyone.

For the foreseeable future, ICPSR expects that much of the data supplied to users will continue to be provided on magnetic tape. All surveys of our users indicate that magnetic tape remains the overwhelming preference as a transmittal medium by the majority of our users. (However, there is a great deal of effort going into determining the next generation of reliable storage and transmittal media.) Additionally a significant number of our collections will not be suitable for transmittal by any other medium for the foreseeable future. This is largely due to the size of many of the collections which span several reels of tape. It is expected that the format of some collections will initially preclude their transmittal on other media. Hierarchical data files will be best supplied on magnetic tape at least for the time being. Nevertheless, the ICPSR has been taking steps to move toward other alternate media.

In February, 1991 nearly 100 copies of a two-volume CD-ROM containing the Panel Study of Income Dynamics data for waves 1-20 were distributed to Official Representatives at member institutions who had expressed a wish to participate in a field test of the product. The data were

supplied on CD-ROM in raw, ASCII format. SPSS/PC and SAS/PC statements were provided on each CD-ROM. Users could use the statements to prepare extracts from the main file or they were free to utilize their own software to perform the extracts. Responses on the questionnaire that was provided with the field-test copies indicated that users overwhelming approved of this approach for the CD-ROM.

ICPSR expects to produce a limited number of CD-ROMs in the future. It is expected that collections selected for this medium will be those that users have indicated they would like to see in this format, and those collections that have a high distribution volume. Some of these additional CD-ROM products should be available within a year.

The Computer Support Group of the ICPSR is in the process of conducting tests with a group of sites to evaluate the transmittal of ICPSR data through Internet. When these tests are concluded and relevant programming that supports this activity completed, we expect that access to ICPSR data will be expanded to include public data networks. It is expected that while eventually all ICPSR data will be accessible through the public data network, initially selected collections will be available in this manner. In order to make ICPSR data available through Internet, the thousands of files in our holdings will need to be moved from exclusively magnetic tape storage to optical disk storage. Since this will be a relatively large task, not all data will be stored on optical media immediately.

Finally, ICPSR is in the process of identifying collections that will additionally be available to users on floppy diskette. Data collections selected for floppy diskette production will be those for which there is demand for the data to be provided on this medium. The data collections that will be provided on diskette will be those that do not require large numbers of diskettes to accommodate a given data file. Additionally they will be collections that are available in raw, ASCII format. It is expected that the data will be supplied with self-extracting compression software with the appropriate README files that provide users with relevant information about the contents of the diskette.

As ICPSR continues with the installation of both software and hardware that will allow us to move from magnetic to optical storage media, we expect to continue to explore the feasibility of adding new services and upgrading older ones. While the ICPSR will monitor and respond to the technical changes many of our users are experiencing, we will also continue to remain responsive to those users who are not experiencing rapid technological changes. For the foreseeable future, ICPSR will seek a balance that allows us to serve users spanning the full spectrum of technical capabilities.

[1] Presented at the IASSIST 91 Conference held in Edmonton, Alberta, Canada. May 14 - 17, 1991.