ON-LINE OR ON TAPE

JUDITH S. ROWE
PRINCETON UNIVERSITY COMPUTER CENTER
PRINCETON, NEW JERSEY

(This paper was delivered at the 1981 IFDO/IASSIST Conference, Grenoble.)

Few of us here remember the early days in which data were transported from one installation to another using metal tapes or punched cards. Needless to say, not many data sets were transported. Metal tapes are now historical artifacts, although most of our computers can still read punched cards. But for almost twenty years data archives and data libraries have used magnetic tapes as the medium of choice for data exchange. With the advent of microcomputers some exchange uses floppy disks, but since no standard has been established for these and since their capacity is quite limited, they are not widely used for this purpose.

To a large extent, access to data by local users is also via magnetic tapes, although under certain circumstances data can be made available permanently or temporarily on disks. The decision to provide local access and certainly to provide permanent storage by one device or another is essentially one of cost and is based largely on the local charging algorithms for the storage and use of the various media. The two key variables are the size of the data set and the frequency of its use.

Although two computer centers may both be attempting to recover costs, they may not necessarily be doing so in the same way. Charges are normally set to encourage particular types of user behavior. Most centers, for example, offer cheaper off-hour rates in order to even out the flow of work. However, depending on the size of the machine and/or storage area, and the number and the type of input-output devices, charging may encourage the use of tape storage or disk storage, high density or low density tapes, permanent or mountable disks.

At Princeton it clearly makes sense -- and will increasingly in the future as we eliminate mountable disks -- to store small, frequently accessed data sets on permanently mounted disks and large, infrequently accessed data sets on high density tapes. However, it should be noted that in order to keep the number of tapes in the machine room manageable, there is a small penalty for archival tapes which are seldom or ever mounted. As the computer center's largest tape owner, the data library owns its own tape rack and hence does not pay the "no use tax" except for rented tapes in the public use area.

With the present array of storage options now available to center users, decisions concerning the middle-sized data set may be less clear. Princeton provides charts to help users pick the most economical storage medium for a given data set. These have now been augmented with a small, publicly available program which allows the user to insert data set size and anticipated use variables in order to obtain a storage medium recommendation. The charts, however, are useful if one is not in a gray area.

Until recently our concerns with these issues were purely local ones, but with the increasing amount of data now available from on-line services, a new dimension has been added. It is no longer enough to think in terms of appropriate media for local storage and use. Many of us are already faced -- and all of us will be increasingly in the future -- with new decisions on the form in which data should be acquired. These decisions are not easy ones, and they will not get easier in the future. There is no single answer which applies to every data archive or data library, nor is there a single answer which applies to every data collection.

We cannot say, e.g., that Princeton should buy tapes and the University of Grenoble should use on-line services, although depending on the local environment some institutions may lean in one direction or the other. Nor can we say that everyone should acquire data set A on tape and data set B through an on-line service, although there, too, we may all veer in one direction or another.

In making these decisions there are three major components for consideration: our local environments, which although essentially fixed are different for each of us; our user communities, which may vary both among us and for each of us over time; and the characteristics of the various data products available to us, essentially the same for each but differing from product to product. It's a little like running a restaurant. We know what our local resources are -- the size of the ovens, the number of tables -- and generally speaking, we can find out what products are on the market and the characteristics and costs associated with acquiring or using them -- fresh, canned and frozen meat, fish and vegetables. The big question is how many people will turn up for dinner.

If we have lots of group reservations made well in advance, we can plan more easily; but if we have only a walk-in clientele, then we have only our experience and the complaints of the customers to rely on. Nonetheless, in spite of these local and market variabilities, it is likely that in the future we will each find ourselves using a mix of "on tape" and "on-line" resources The remainder of this presentation focuses on the elements which will determine your particular mix and mine.

Let us first look further at local computing environments. Although even within a given institution some of these elements may vary for different classes of users, by and large there is a pattern which characterizes each environment.

The first and perhaps the most critical element in this pattern is the type and the amount of money available. In most computing environments, we find two types of money. The first is real money, money which can be spent as easily outside the institution as inside, money to which there are no strings attached.

The second type of money is internal -- general funds or "funny money" -- money which may be available only for local computing (or, in some cases, for other local services) but which cannot be spent outside. In each institution both the data archive or data library and the individual user have some combination of these moneys. The total amount of each and the amount of computing and ancillary services which the internal component can purchase are key in the on-line or on tape decision. If there is little real money, but adequate

internal money and cheap computing, the inclination will be to keep it local, to buy tapes.

Assuming, however, that the money issue is not so elastic, what are the other elements in the local environment which will affect our decisions? Certainly the quality of computing is one. Although the reliability of the system, the cost and availability of storage devices, and turnaround or response time are important, the quality and variety of available software and programming support are key. If it is not possible for the user to do locally the kinds of analyses required for his work, and on-line services to do these things are available, the pressure to use them will be great.

Last but not least is the quality of the data archive or data library staff and the systems available for acquiring, storing, locating and accessing data. If these are all well-organized, ceteris paribus, users may be less interested in paying real money and learning new systems to use on-line services. Other considerations peculiar to each local environment may also be at issue, but those I have mentioned are always relevant.

Now what of the individual data products? The most fundamental question is: Is there a choice? Some data are available only on-line and other data only on tape. One may characterize data files as consumable, durable, or permanent. The consumable files are like food. One must replace them daily, weekly or monthly. These are files, e.g., which provide input to economic models where the presence of timely data is essential. Timeliness is generally far more essential in business than in academia. These data are generally available only on-line, which accounts in part for the fact that 90 percent of the use of on-line services for accessing numeric data is by business and only 10 percent by academia. For the rest of the explanation we must work back to differences in local environments, to the differences in how commercial and academic users measure costs.

Durable files are like cars, refrigerators, or even husbands. They may be inert files or dynamic files in which the currency of the models is valuable but not essential. The availability of these files on-line is usually a function of size and generality. Large, small-area aggregate files or special-purpose microdata files are less likely to be available on-line than files of national, state, or provincial annual time series data.

Permanent files are like fine paintings or Oriental rugs. They have long lives and small groups of devoted admirers, but seldom enough admirers to justify their availability on-line, whereas their long lives make them good investments for purchase on tape.

Assuming one has a choice between on-line or tape access, what other characteristics of the data product should be explored? Size, cost, and subject matter are certainly factors, as are the relative ease and appropriateness of access and use -- i.e., the source of the data, the form in which they are provided, and the availability of software.

If the file is small and inexpensive, can be purchased and delivered within the appropriate time, can be analyzed locally, is of a permanent nature and relatively general interest, buy tapes.

This decision, however, will also be affected by expected use. For exam-

ple, during the life of the file will there be many users or few? Will they require access to the entire file or only to a small subset? If the latter, will they each require a different subset? And what will be the nature of their use? Do they require only a one-time display of data or will they be subjecting the data to complex and repeated analyses? Can one purchase only a subset of the data on tape or is it necessary to purchase the whole collection? Is there a subscription fee for on-line access, or does one pay per use?

If the cost of on-line access is moderate and the user needs to display only a few numbers, then purchasing the tape may not be justified unless there are clear indications of high future use.

I have identified many variables in the decision equation which we may someday construct. However, our experience is as yet too limited to assign values to those variables. At this point we can only make subjective judgments, based on our awareness of the relevant factors. Each of us behaving rationally may make different decisions, because our computing environments and our user communities are different. Moreover, the advent of cheaper, larger on-line storage makes these decisions dynamic ones.

Some data we may all acquire on tape. Other data we may all access on-line. But for an increasing amount of data our decisions will be individual ones. There is no single answer.

A detailed look at a few specific examples of available data which some of us have acquired or might consider acquiring may serve to illustrate the interaction of local environment, user community, and data product characteristics in determining data access choices. I have tried to choose an international group for these illustrations, but you will forgive me if I speak more about the data I know best.

The United States Bureau of Labor Statistics (BLS) maintains a collection of 100,000 time series covering such subjects as national labor turnover; industry, producer, and consumer price indices; and national, state, and area employment. This collection is known as LABSTAT, and in BLS parlance LABSTAT includes both the data and the software for its analysis. Many of us over the years have purchased individual time series or groups of time series from the Bureau for the cost of approximately $100 per reel, and have used locally installed time series packages such as TROLL, TSP or SAS/ETS for data analysis. A few of us have used one or more of the selected series which a number of the on-line services make available with accompanying analytic software.

Recently Lockheed, which specializes in bibliographic data files, has made available to its users almost half of the LABSTAT data collection. The primary omission is the series relating to unemployment. Lockheed acquires each new update as soon as it is released and provides access to it, using the software which is familiar to its bibliographic data file users as DIALOG. There is no provision for any manipulation of the data, but selected displays are simple and inexpensive.

As we discovered at Princeton in providing access to the United States decennial census data, there is a large group of individuals who use machine-readable data products to find numbers which are not available in print or microform or which are more easily accessible in machine-readable form.

Lockheed has made the same discovery, and the "Labor Statistics (LABSTAT)" file is already being heavily used. A United States Bureau of the Census file called "U.S. Exports" was added at the same time; and it is likely that Lockheed and its competitors will soon add similar data files to their available holdings, and that they will concentrate on the display market rather than on the analysis market. For those of us who wish access to the full data base and to analytic capabilities, it is still necessary to purchase tapes.

The International Monetary Fund produces several major data collections, including the World Financial Series (WFS) and the Direction of Trade. The former contains over 40,000 time series of financial statistics on over 160 countries, plus aggregate data for the world and over 50 regions. These data go back to 1948 and have been updated monthly since 1965.

The Direction of Trade contains approximately 100,000 time series of import and export statistics for 230 nations and their trading partners. These data also go back to 1948 but have been updated monthly only since 1977.

Monthly tapes for IFS may be purchased directly from IMF for an annual subscription of $400. ICPSR members may obtain both collections on a less timely basis without cost. Regardless of the source, when the time series package arrives it is necessary to use a special COBOL program to unpack it and write it on disk, an assembly program to reformat it to tape, and a FORTRAN program to retrieve it for analysis before using the data. Data Resources Inc. (DRI), possibly the largest of the commercial on-line sources of economic and other statistical data, ADP Network Services Inc., FRI Information Services Limited, Rapidata Inc., and Telesystemes-Eurodial, among others, provide subscribers with access to these data and to software for analyzing them.

At Princeton the International Finance Section of the Economics department has adequate general funds money for computing and a graduate assistant to preprocess the tapes and to analyze each month's data. As a result, on-line offerings have not proved attractive.

DRI also provides United States Current Population Survey annual time series data from 1968 to the present. These time series are based on the aggregate data published by the Bureau of the Census. For individuals interested in CPS data in this form, there is no competing product. However, for individuals interested in cross-sectional microdata and specifically in the monthly supplement questions, the Bureau sells each month's file for under $300. The quality of the files themselves and of their accompanying documentation has improved markedly in recent years, and normally the data are easily analyzed if software for processing hierarchical files is locally available. The popular March file, known as the Annual Demographic File, is available from ICPSR.

A somewhat unique case is the 1970 United States Census of Agriculture. A preliminary tape containing a subset of the preliminary published aggregate data is sold by the Bureau at its customary $110 per reel for 5 reels. At the request of some members of the user community, final data were produced on tape as a special tabulation and sold for $1000. In the most recent Directory of Online Data Bases, published by Cuadra Associates, an organization called On Line Research was reported to be making Census of Agriculture data available on-line. Further investigation, however, uncovered the fact that this organization is no longer in business. In this instance it would appear that the 'display-only' customer will do well using the printed reports, and that

the researcher whose data needs are not excessive may find it more economical to reenter the necessary data himself -- particularly if student or clerical time is freely available.

This latter alternative is not one I would normally encourage, but it should not be totally overlooked. It may sometimes prove to be the most practical alternative.

In discussing on-line access I have not distinguished between commercial vendors and academic sources. The issues are essentially the same. When should data be acquired and used locally, and when should it be accessed on-line from a remote site? We must each make our own decisions, and perhaps when we gather together next time, we will have a larger body of experiences to share.

I realize that many of you had hoped for more specific recommendations on these issues. However, as with so many issues concerning the management of data archives and data libraries, we are each bound by our unique institutional settings and by the peculiar nature of our user communities. In the case of our local environments we must consider the amount and the type of money available for computing, the cost and the quality of local computing, the software available and supported, the amount of cheap labor, and the nature of our own operations. In terms of the user community we must consider the number of anticipated users and uses, the type of use, and the amount of data required. And finally, in looking at each data collection we must be aware of the options from which we may choose: the size of the file, its durability, its content, its completeness, its currency, and the software required and available for on-line or on tape use.