Downloading for PC Users; Part I: The U.S. Government Experience

by Donald F. Harrison
W. Jon Heddesheimer
National Archives and Records Administration[1]

summary

What follows is the edited manuscript of a dialog presented at the Marina Del Rey meeting of the International Association for Social Science Information Service and Technology, on May 22, 1986. It focuses on the programs of four Federal agencies which download onto flexible diskettes information traditionally offered on tape. After describing various ways of writing data on diskettes, formatting considerations (DBMS or ASCII) are discussed. Past and future changes in the cost, speed and capacity of micros are stressed. Program details such as pricing and contractor versus in-house production of floppies are mentioned. The point is emphasized that Federal downloading is restricted almost entirely to small, simple files formatted in a DBMS because agencies do not consider the flexible diskette suitable for storing large amounts of data. Major statistical producers expect mass storage devices like the CD-ROM to become generally available to micro users. They further anticipate that soon the average PC user will be able to download and manipulate large ASCII files. The authors discuss various electronic data communication systems, such as the Navy's DIF and ISO 8211 as alternatives to media transfer. This leads finally to a discussion of digital communication methods such as BITNET, NETNORTH and EARN. The authors conclude with predictions of the effect on the future operations of data archives.

•

Jon (Introduction): In this session we intend to describe the efforts of four Federal agencies to serve PC users by downloading onto flexible diskettes information traditionally offered and sold on tape. We will use a dialog format to encourage you to interject your own comments.

Don: There are many different ways to define the term downloading. My IBM PC Users Manual defines it in terms of simply "printing out" material in hardcopy. My PC TALK Users Guide defines it as receiving data from a remote terminal. For the purposes of this session, the "down" part of the word refers to moving data "down" from a mainframe (or mini) to a smaller computer (mini or micro, in this case to a micro), and the "load" part of the word will refer to the writing of that data on a 5 1/4" flexible diskette — commonly referred to as a "floppy" — so that the data can be

[1] The opinions expressed in this article are solely those of the authors and in no way reflect the official position of the U. S. National Archives and Records Administration.

manipulated on a microcomputer. In this session we will concentrate on downloading as a reference service to the user public.

Jon: The following four possibilities by which automated records can be downloaded from mainframe files onto floppy diskettes for users of microcomputers have been suggested in our research:

a. Downloading files in bulk. Using a package, the 'downloader' transfers an entire file in straight ASCII. This requires careful calculation to make certain the file does not overwhelm the PC's capacity. Also, because the data are relatively unformatted, the PC user will need to be well acquainted with the mainframe's procedures of access in order to retrieve and use the data.

b. Downloading files in a data base management system 8DBMS) format, such as LOTUS or dBase. These programs pre–format the information to allow the user to begin work at once. The vast majority of floppies offered by government agencies are in a DBMS format.

c. Selective data access. This allows the user to request selected portions of mainframe data either from a single file or across a range of files tied together in some fashion. This can be accomplished using generalized menus or customized packages, such as, in a corporation where the command "marketing department budget" might trigger the downloading of abstracted relevant information.

d. Cooperative processing. This method involves expensive custom packages which allow the user to establish an intelligent connection between mainframe and microcomputer applications.

Don: Agencies and institutions which make machine–readable data available to researchers find that there is a growing interest in storing and manipulating data by microcomputer. Some researchers are reluctant to use traditional computer service centers with mainframes and programmers. They would rather perform the research themselves in the privacy and convenience of the office or home. Moreover, today's microcomputers have the storage capacity and the sophistication of yesterday's mainframes.

Jon: Yes, Don. This is part of a general trend away from large mainframes and central processing facilities, accompanied by a great increase in the capacity of PC's. Offering data on floppies is one way for an archives to harness this trend and increase visibility and clientele.

Don: In preparing for this session, Jon and I located several very small Federal offices which download for their users on a "swap" basis: users send in formatted flexible diskettes which the agency writes on and returns to the researcher. The National Register of Historic Places, maintained on–line by the National Park Service, is one example of this practice.

Don: This presentation, however, will concentrate exclusively on the experiences with downloading of four Federal agencies. These agencies have had considerable experience distributing data on tape and, just in the past few years, have begun to write data onto diskettes. They are: the National Technical Information Service (NTIS), the Bureau of Economic Analysis (BEA), the Bureau of Labor Statistics (BLS), and the Bureau of the Census (Census).

Don: NTIS was created by an Act of Congress to make available, to the public, material created by a variety of Federal agencies because these agencies do not have revolving funds (such as the National Archives Trust Fund) with which they can sell copies of records to the public. Over the years, NTIS has become a broker for both textual and non–textual records.

NTIS's revolving fund allows it to set prices, receive monies, and publish catalogs. NTIS sells its technical information products and services under the provisions of Title 15 of the United States Code. In the early 1970's, it started to accept machine-readable data files (MRDF's) from Federal agencies and, since then, has been selling copies of tape files to the public. In the last year or so, NTIS began a program whereby any of its MRDF's could be ordered on flexible diskettes written in straight ASCII or in one of several commercial, packaged data base management systems. NTIS sees absolutely no problem with very large orders, and it is up to the customer to determine if her/his microcomputer can accept the file volume; one popular offering is written on 87 floppies. NTIS is the largest seller of downloaded data; it far exceeds all the others combined.

Jon: BLS was established in the early twentieth century and has been in the forefront of statistical analysis from the beginning. BLS, as well as the remaining two agencies to be discussed, believes in making its files available to the public directly from the analysts who created them. Consequently, BLS has for 15 years been publishing a catalog of data files available on tape. Two years ago it launched a program of making them available on flexible diskette as well. Each division within BLS does its own analysis, its own downloading, and sets its own prices.

Jon: BLS is uniquely qualified to download data due to the existence of a LABSTAT "umbrella" system. Created for in-house research, over 25,000 search data elements can be tapped across an enormous range of statistical data. Because this agency thus has the capacity to allow researchers to browse through the data (on-line) and pick and choose from all files, it has the capability to pioneer a similar downloading service to the public.

Jon: The BEA, like the BLS, believes that its analysts alone are capable of dealing intelligently with the user. Therefore, it makes data available to the public directly, not through any broker or archives. But its files are so voluminous and complicated that the downloading program encompasses only one data set, which was previously offered on microfiche. Therefore, this modest program is operated "in-house" by one person, merely as an extension of an existing service. The agency has no plans to expand the program.

Don: Census has been gathering statistical data since 1790, and began working with machine manipulated data with the 1890 decennial census. Not unexpectedly, Census has also been in the "downloading business" longer than any other Federal agency. The Data User Services Division, which has been making public use sample data files available on tape for many years, began in 1984 to make these same files available on floppies.

Jon: Yes, this program was the first. All agencies contemplating using downloading as a part of their reference service began by visiting Census. The agency spent a great deal of time and money creating custom packages which subdivide extensive databases into "PC-sized" chunks.

Jon: Thus we are discussing four programs, the oldest of which first offered downloaded files in March, 1984. This is a new service for the Federal government, which has thus far been slow to provide data on floppies. This is due partly to lack of resources for new services. But even more important, it is our contention that the low level of downloading activity is due to beliefs about the immediate future of the PC and the data medium it uses, which we will discuss in this dialog.

Don: One must calculate carefully before writing mainframe data onto floppies. In the first place, not all files can be downloaded to the PC because of manufacturers' limitations and specifications of the diskette. For example,

one must consider the size of the file as it exists on the mainframe and the complexity of the language in which it is written. One flexible diskette, filled to capacity with a flat ASCII file, contains 362,496 bytes, (DOS 2.1 formatted, double sided, double density). This figure is insignificant when compared to a standard reel of magnetic tape (11" reel, wound with 1/2" tape, 2400' long, and filled to capacity with 9 track, 6250 bytes-per-inch). Such a tape could contain approximately 120 million bytes, or the equivalent of more than 150 floppies. Furthermore, when the data are formatted on the diskette to accommodate a microcomputer data base management system such as LOTUS or dBASE, the diskette will probably hold far less than 362,496 bytes. Of course, the obvious solution is to add more diskettes. NTIS offers one cartographic data file from the Central Intelligence Agency called "World Data Base II" written on move than 80 diskettes. In order to manipulate the entire file, one would have to load all 80 diskettes into the PC first. None of the other three agencies offers files of this magnitude on floppies, because they believe that users cannot or will not use them.

Jon: To illustrate Don's point, one can subdivide downloading programs into "active" and "passive." One agency makes everything available from its extensive holdings through a contractor and in most commercial DBMS formats. If the researcher is comfortable with 200 floppies, so be it. This is "passive" and a growing trend. Another agency tries to be somewhat active by limiting what it will offer to 2 diskettes per file, but does little beyond that. Two agencies offer their holdings only in LOTUS 1-2-3 format. One agency, by far the most "active", spends a great deal of time and effort subdividing complex files into compact units convenient to the PC user.

Jon: Outside the Federal experience, but certainly worth mentioning, is one data archives with limited resources which offers workshops on how to download and format data for use on a PC. This approach encourages users to master a FORTRAN program which enables them to download, and thus make greater use of the tapes held by this archives. They are saying, in effect, "We will give you the tools and turn you loose." Of course as a PC owner, you have to be very serious to use this as a research strategy.

Don: Software formatting is a second consideration. Many persons we interviewed agreed that all micro users are divided into two types: the mainframe expert "data junkie" who can work with unformatted "flat" data, and a newer, less energetic user on the scene. This second user prefers to insert a floppy, flip a switch, and let the machine do the rest. In many cases, this new user just purchased her/his PC last week and expects instant results. At least two of the four agencies we spoke with accommodate this latter type by formatting the data in a data base management system which allows the user to begin data manipulation immediately. (LOTUS and dBase are two of the several DBMSs available). Formatting involves considerable reworking by the data producer. The payoff, however, is in attracting many more users, in fact an entirely new market of users, far different and more numerous than those who have traditionally ordered magnetic tape for use on a mainframe. We suggest that these clients are best served with data distributed according to the specifications of the individual order. This avoids consumer complaints and unpleasant scenes with commercial software manufacturers.

Don: Hardware formatting is another matter completely. There seems to be a universal preference (among data producers, data brokers and data users) for IBM-compatible diskettes. Most prefer DOS formatting over CP/M as well, though this doesn't seem to be an issue.

Jon: I agree that using hardware designed to be IBM-compatible is far more important than

deciding on one software package. I would like to add that in the next few years, the ordinary user should be able to deal with software independent information and do easily what today only a "data junkie" can accomplish. First will come graphics packages, a process already underway. Next will come the tools for the inexperienced to handle large quantities of raw data. Along these lines, I expect the storage capacity of micros to increase two-fold in two years and ten-fold in five, with little increase in cost. Software designers expect this to happen and already are hard at work.

Don: The strategies of pricing suggest as many solutions as there are agencies. They depend on what the individual agency perceives its internal costs to be, whether or not it is dedicated to the concept of service to users and, in the case of agencies using a contractor, a markup from the contractor's costs. These strategies may be compared with the production of microfilm publications in traditional archives and libraries. Data producers (tape or floppy), like microfilm producers, may produce the product on demand, and charge the first customer the full amount required to recover costs; the second and third customers, in turn, pay only a marginal fee. A second solution is to spread the charge over several users. Yet another solution is to absorb the cost of preparation and simply charge a flat fee. In addition to production of microfiche publications, this range of solutions seems to occur with production of files written on either tape or diskette. In the case of diskettes, there is great diversity in charges: one agency charges $35 per diskette; another charges $75 for the first floppy, and $15 for each additional diskette in the same file; yet another agency now charges $60 for the first diskette and $12 for successive ones. Our fourth respondent, by producing only one file for the public and updating it each month, charges a flat $240 per annual subscription of 12 monthly installments; this amounts to $20 per diskette.

Jon: Circular A130, issued in 1986 by the Office of Management and Budget, instructs agencies to charge incremental costs incurred in serving researchers. Information is defined as a marketable resource, thus tacitly refuting the notion of public service. Pricing policies vary. For normal orders, the charges expressed are "so much per diskette," not "so much per file." In general, large user service organizations charge handsomely for special considerations, special formats, special tabulations, etc. One agency representative stated that "by-the-book" processing charges are so extreme that the final output (tape, floppies, hardcopy) costs about the same regardless of medium. This is partly due to the cost of getting the data ready for the user. Processing data so that the user can work with them frequently constitutes the major portion of the entire cost of a service order. And, of course, a part of this problem is also that to write data onto a floppy sometimes requires reworking and reformatting.

Don: Since floppies are fragile, hold a comparatively small amount of data and can be accidentally erased, each of the four agencies we spoke to have considered other media for use on the PC. Almost all our contacts discussed the use of the compact disk for digital data storage. Since the data are written with a laser, there can be no problem with accidental erasure or overwriting. It is a read-only mode. Unfortunately, it requires a separate and extremely expensive disk drive. All persons we spoke to predicted that compact disks will soon be both plentiful and economical. What makes the "compact disk-read only memory" (CD-ROM) so attractive is that it will store the equivalent of 1500 flexible diskettes — or the equivalent of 4 high density mainframe tapes.

Jon: Data professionals are also giving consideration to an interactive storage device with the characteristics of hard disks. This, if it becomes a reality, is further down the road than CD-ROM. (At present, "read only" is considered a virtue.) The Bernoulli Box, while

too expensive for individual use is nonetheless indicative of what might become commonplace once PC's are better able to accommodate and manipulate large data bases. Thus one can confidently predict that a standardized, economical, mass storage replacement for floppies will soon be available. The average user will not rely exclusively on floppies and indeed, may not use them at all.

Jon: However, considerations might be reversed in the future when pondering whether or not to store data on the CD-ROM or the Bernoulli Box. There might be a case where too little data is requested, even for the economical use of a diskette. Such cases are tailor-made for the electronic bulletin board, which is designed to transfer small bits and pieces of data rather than huge data bases. One agency routinely gives small bits of data to users for the cost of a phone call instead of charging the price of a floppy. Electronic bulletin boards are becoming increasingly popular for the presentation of finding aids, lists and other advertisements.

Don: Electronic bulletin boards depend on electronic data transfer. Instead of writing the data onto a diskette and shipping the diskette to the researcher, it involves sending the data across a telephone or wireless circuit from the archives in which the data are stored directly to the researcher's microcomputer. It is a method that is gaining in popularity and could easily be the preferred data transfer method by the year 2000.

Don: However, electronic data transfer also introduces problems of interchange formats. These involve the use of data filtering devices which are important when machines of differing specifications from several manufacturers and using different software are communicating with each other on-line. In March, 1983 the U.S. Department of the Navy initiated a cooperative effort among government and leading office systems to define and test a Document Interchange Format (DIF) which vendors could support. Today DIF permits the interchange of textual data between word processors, providing about 95% of their document formatting needs. Twelve of the fourteen manufacturers who cooperated with the DIF test were Datapoint, Data General, DEC, Hewlett Packard, National Cash Register, Sperry, Motorola, AT&T, Four-Phase, Xerox, Wang and IBM. The DIF generally would filter textual data files between microcomputers. But what about statistical data files? And suppose one of the computers is a mainframe communicating with a microcomputer?

Don: At about the same time that the Navy developed its DIF, the International Organization for Standardization developed a set of standards which incorporates a mechanism allowing statistical as well as textual data structures to be easily moved from one computer system to another, independent of the manufacturers. The resulting system is called, "ISO 8211." It is much more flexible than the DIF and will accommodate magnetic tape, disk packs, flexible diskettes and data interchange over communication lines —in any combination, either as a source or a target. It will accommodate files with variable length records as well as those with fixed length records. User file structures such as sequential, hierarchical, relational or indexed, could be connected with the interchange structure. Therefore, it can be said that ISO 8211 is both content and media independent.

Jon: Once an operation involves more then a few files, or even perhaps from the very beginning, employing a good contractor is probably the best solution to the problems of disseminating downloaded data. The risks of floppies becoming obsolete, or of incorrectly anticipating researchers' needs are passed on to the contractor. Contractors with extensive experience are numerous, and will allow the researcher (for a price) to define his specifications. Contractors now offer agencies a flat price of under $25 per floppy, even if they

have to copy and reformat the original tape, and for this price will keep a copy in a contractor maintained library. Thus it is now possible to make money on an initial order and still charge reasonable prices, a situation only true in the last few months. Before that, even the largest organizations found it necessary to sell several sets of a file before recovering their costs.

Don: An undeniable advantage of an in-house operation, especially in the early stages of building a floppy program, is that you can work with a researcher in developing files. You can also help her/him select a simple, established file to get the "feel" of things. Individual program managers still tend to offer this flexibility in decentralized agencies (in two of the four agencies we contacted). Also, an agency can send new files to sophisticated, established users for comment prior to public release. There will come a time, however, when economics coupled with instructions from the Office of Management and Budget, will make such operations impossible for all four agencies we contacted. The present reality is that these four agencies will probably have the choice of using a contractor or having no program at all.

Jon: The justification for modest or non-existent downloading programs in major statistical agencies is that floppies will soon cease to be the medium of choice for downloading to PC's.

Don: Also, given the consensus of opinion that micros will vastly increase in capacity and that the ordinary data user will be able to deal successfully with unformatted "flat" ASCII data, it pays to look down the road rather than be frozen in the present.

Jon: I recommend that you use electronic bulletin boards, first as a catalog to advertise holdings and later to hold simple updates and smaller data bases. Using an electronic bulletin board as a catalog and as a vehicle to answer routine researcher inquiries should greatly increase the visibilty of your collection while eliminating a great deal of your reference load. Ideally, reference personnel should deal only with special requests. To ask reference personnel to answer the same questions over and over is both expensive for administrators and demeaning to professionals. GTE SPRINT's "PC Pursuit" represents the wave of the future and is an example of how the general public can tap into electronic bulletin boards, or even on-line data bases. This relatively new service offers unlimited nighttime and weekend access to 14 cities from over 200 TELENET areas for $30 per month.

Jon: I would suggest also that you consider designing "umbrella" systems which tie your holdings together via common data elements, making the researcher her/his own boss regarding data access. I have already mentioned LABSTAT. Another example is to be found in the networks which have developed between IBM installations throughout the world. Started originally as independent data collections to service specific needs, they have gradually become integrated and can be accessed by the casual user via standard menus, (given, of course, the constraints of access levels.) Unlike LABSTAT, which was developed from the beginning to be somewhat like it is now, these systems were cobbled together artificially to fend off competition from other sources and to satisfy internal corporate requirements. They serve as an excellent example of how data archivists can develop menus to link together their own collections and eventually develop the means whereby researchers may consult numerous data archives utilizing standard search commands.

Don: This technique is already in use in libraries and manuscript collections in the United States. Consortia of library collections are accessible by computer networks in at least two very large systems: the Research Libraries Information Network (RLIN) and the On-Line Computer Library Center (OCLC). However,

for the most part, what one reaches via these systems are descriptions of the collections, not the collections themselves. This is very different from our thesis. What we are advocating is that the data, as well as descriptions of the data, be made available to remote locations, in a library research area, for example, or in a researcher's work/terminal area, in short, anywhere a researcher has access to a microcomputer and a modem.

Don: Variations of this are at work in digital communications networks. Collaboration between academics at widely separated locations is becoming more and more common as technology allows them to exchange ideas and information more easily. At research locations throughout the United States, a network called BITNET is playing a major role in this information interchange. BITNET is a cooperative digital communications network connecting over 1,200 computers in universities and other educational and research institutions. By connecting to BITNET one can gain access to computers on the MAILNET, EARN (Europe) and NETNORTH (Canada) international networks. Using the facilities at any one of these four systems, data archivists could easily transmit part or all of their holdings to researchers located within the geographic confines of the others. A combination of the filtering systems laid between hardware of diverse manufacture together with the use of the new, world-wide, inter-computer communication networks will revolutionize the transfer of electronic data.

Jon: Data archivists, therefore must ponder the integration of their own holdings using umbrella systems, while learning to relate these to the contents of other repositories through digital communication networks. While accomplishing this may initially require additional resources, long run costs should decline or stabilize while user access increases.

Conclusions

As data archivists we must observe this research trend of exchanging mainframes for micros in order to perform statistical and other kinds of analysis on machine-readable records. A reference program offering data on floppies through a contractor is a good way to begin tapping this market which is already far larger than that for tape. Further advances in the capacity of micros coupled with new software, an inexpensive mass storage medium to replace floppies, and continually declining costs will allow individual PC users to function more and more like data centers. All this should only serve to increase the market for downloaded machine-readable data files.

Information transfer via floppy diskettes or other media is but one way to move data from mainframe to micro. An electronic bulletin board could be started by a catalog listing, and later enlarged using electronic data interchange over one of many communications networks such as BITNET.

Finally, archivists should plan to develop "umbrella" systems which will tie holdings together in a given repository and ultimately establish access to other collections.□