CD-ROM and the data archive:
beyond retrieval

by Ann Gerken[1]
University of California, Berkeley

Optical storage offers data archives a new medium for storing data and a change in the information retrieval environment. As with any new technological development, it is important to evaluate the impact and tradeoffs of adapting to new equipment and systems. By stepping back for an overview of information storage and access, better evaluations may be made within the context of the "big picture" of data use. Before embracing this new technology, it is important to understand where we are going and how optical storage can help us. Optical media offer some solutions to storage problems, but have not yet shown themselves to be solutions to the overall needs of information providers and users.

First of all, thanks go to Forrest Williams, Census Bureau, and to George Hall and Courtney Slater, of Slater-Hall Information Products, and to Ed Spar, of Market Statistics. These people have provided us with CDROM test disks for our evaluation, and I sincerely appreciate their support. Some of the points I will discuss today have been included in the report of the State Data Center's Subcommittee on CDROM. This report is available from John Kavalaunis of the Data User Services Division. My ideas do not necessarily reflect all the opinions of that subcommittee.

Consider the advantages of optical media as storage devices. CDROM and WORM disks provide a minimum of 10 years' secure storage. The disks provide random access, and are portable and compact. However, there are some problems. For most data archives who rely upon large centralized computing systems, the shift from storing data on tapes and disk packs to storing data on microcomputer peripherals means a dramatic shift in responsibility. How will data archives provide multiple user access to data on microcomputers? Can archives handle the equipment and network burden? Another problem with CDROM is that archives cannot produce CDROM disk copies in house; there is no local recording capability. This

[1] Presented to The Association of Public Data Users October 22, 1987, Washington, DC

means that if data archives are to produce optical backups of their holdings, they must use WORM disks, and will have to acquire and maintain two kinds of optical devices. Other problems are access speed, and in some cases, limited storage capability (CDROMS only hold 4 reels of tape, and STF4 for California takes up 14 reels of tape). Another issue to consider is that the software to access the data on CDROMs and WORMs has not kept pace with the technology. Even though information on microcomputers appears to be closer to the end user, the interfaces and support for access are often limited. And finally, we must ask when will optical storage be cost effective, and when will the major data distributors begin to use this technology?

Information retrieval is the process of extracting specific data items from specific records in a data file. At this time, information retrieval software is produced for each CDROM product since standard retrieval software for statistical data files is not yet available. At its best, this retrieval process is informed, flexible, logical, and fast. However, since each data file comes with its own custom software there is a wide range in the quality of these software interfaces. Users will be required to learn multiple retrieval processes. Also, statistical CDROM products are more expensive to produce since producing custom software for each data file requires extensive programming.

The premastering and production of the optical disks themselves are different from producing magnetic tape. The physical process itself is more expensive, and the nature of the files on the disk differ. Since the medium is randomly accessed, "disk geography," or where on the disk various records will be placed, is more important than with tapes. Disk geography can improve performance if done carefully. Disk indexes should be produced and mastered with the data file if the disk is to prove efficient in sophisticated retrieval applications. Error detection and correction codes must also be provided. In summary, "raw" data files on CDROM should be more carefully developed than those provided on magnetic tape. To exploit the technology and access software, optical products should be distributed with value added files.

After information is retrieved, "beyond retrieval", lies the world of post processing where the information is manipulated and packaged to user specifications. At its best, this processing should be supported by an integrated system providing clear steps for users to follow. Information retrieved from CDROM products should be compatible with inhouse software, and transfer of the information should be standard and simple.

In our evaluations, we considered three CDROM products in relation to data product preparation, information retrieval, and post processing. The three CDROM products we evaluated were: Census Test Disk #1, which includes data and software for ZIP codes, population estimates, and the 1982 Census of Agriculture; Slater Hall's 1982 Census of Agriculture disk; and "Your Marketing Consultant" form Market Statistics. The Census disk offers simple extract programs that demonstrate the usefulness of optical storage. The procedures for retrieval are straight forward and can be done by any novice computer user. A highlight of the Slater Hall software is its online help for each variable in the data base, as well as its numeric search capabilities. "Your Market Consultant" offers the capability of sorting data by user defined criteria, and allows users to add their own data to the tables.

Our evaluation at the State Data Program emphasized the retrieval and post processing capabilities of these products, but did not take a close look at the premastering and production of the CDROMs. The issues of disk geography and indexing are yet to be evaluated. However, in the area of retrieval, it is clear that a standard procedure has emerged among these products. Users progress through a series of simple procedures, first selecting geographic areas, then identifying a table or variables, then extracting data from the disk, and finally producing display, report or file output.

What more do archives and users need from this kind of product? Given that a CDROM product provides the information that a user wants, a separate procedure is often required after retrieving information. A user must sort and create custom tables, graphics, or spread sheets in another system. Obviously, there is a need to coordinate software compatibility and command language among in-house software, so that users are not faced with a confusing array of software and transfer methods. In addition, the lack of numeric manipulation and statistical analysis in the CDROM software requires an interface with statistical software currently in use. Data as well as textual information must be moved in efficient ways and in compatible formats.

None of the current CDROM products contain microdata, i.e. data for individual households, persons or establishments. The data that are available are summaries for geographic units, useful for many applications, but are only part of the larger universe of data. Extracts of microdata files are regularly produced in the archive environment, and products and procedures using CDROM technology should be considered.

Our suggestions are that 1) data distributors keep an open mind about which kind of optical media to adopt, so that archives are not forced to develop dual systems, one for data acquisition (CDROM) and one for data file backup (WORM). We also are concerned about the technology becoming outdated. 2) Optical products should be mastered to allow for optimal retrieval by a variety of software, including relational database management software. 3) Retrieval software should provide output that is compatible with other software, and the transfer of data should be an integral part of the software product. 4) Data distributors should view their products not only as stand-alone retrieval systems, but as pieces of a larger information support system.

Some questions for the future: How can optical storage be integrated into the existing information fabric? Who will build the interfaces for informed transfer of data from system to system? Do we place responsibility upon the private sector to develop retrieval products, and how involved can we become? What information systems are in use that identify, retrieve, transfer, and analyze data? How may these systems influence the development of CDROM and WORM products?

Data archives have a long history of responding to technological changes: from IBM cards to low density tapes, from floppy diskettes and random access disk packs to optical media. Eric Tanenbaum wrote: "Archives have a privileged role among information providers for they were among the first to computerize. Thus they offer a rare perspective from which to view the changes." (IASSIST Quarterly, Spring 1986) I believe that archivists will be studying these changes and will offer a unique perspective in emphasizing the integration of new technology into existing services and procedures. Archivists and other information professionals have an opportunity to participate in the development of standard, reliable storage media, to assist in the design and evaluation of powerful flexible retrieval software, and to help create dynamic post-retrieval environments.□