An Analysis Of Cd-ROM as a Long Term Archiving Solution

by Denis Oudard[1]
Digipress

Before we go into this analysis, let's take a look at the history of archiving. This table shows different systems of communication and the role of each element as well as their similarities.

Data	Medium	Coding
Hieroglyph	Stone/Papyrus	Rosetta Stone
Text and B/W Image	Microforms	Language, Alphabet, Magnifying Glass
Databases, Raster Images, ASCII	9 Track Tapes	Tape Drive, Computer System
Sound, Digital and	Compact Disc	CD-Player

All of this combined and/or further processed provides us with information. There is no doubt that at the eve of the information age, we will be called upon to archive, for the long term, huge amounts of information.

In my first paper on this subject I stated that to provide a long term archiving solution, one needs to have two elements:

1. A retrieval or access system that will also endure the test of time.
2. A long lasting medium on which the data is stored.

One without the other is as useless as a deck of cards without aces.

Let's first tackle the 1st objective: access system longevity. Simplicity is the best way to insure the longevity of an access system. Such is the case for microforms. The magnifying glass is a simple system and we can rest assured that humanity will know how to use a magnifying glass for years to come. Therefore, the survival of this access system is virtually guaranteed.

Unfortunately, when dealing with computer archives, simplicity is definitely not part of the equation, so we have to look elsewhere for longevity factors.

The first one is momentum. In other words "How much acceptance or wide spread use does the technology have?" The law of large numbers is going to be a key ingredient. Consumer products have more momentum than professional products because they are sold in the 100s of millions rather than in the tens or hundreds of thousands. CD-ROM is the first computer media to be based on a widespread consumer item. Therefore, we have 100s of millions of machines out there that contain 90% of the key components of a CD-ROM drive. Even if CD-ROM technology, as we know it today, is abandoned in twenty years or so, chances are we will find working CD-ROM drives in a 100 years.

Another important way momentum can be measured is by the number of manufacturers which build the same product. Today I can give you the names of at least 10 manufacturers of ISO 9660 CD-ROM drives. Are you ready? Here we go: Chinon, Digital Equipment, Hewlett Packard, Hitachi, NEC, Philips (LSMI), Pioneer, Sony, Texel, Toshiba; and I know I have left out some.

Compared to the momentum behind CD-ROM, other media are very fragile. WORM drives, for example, depend on the whims of a single manufacturer. Each manufacturer makes a different type of WORM, and the day the manufacturer stops making that drive, you have less than five years to transfer or lose your data.

The second critical factor relating to long-term access is system independence. The reason this is important is because none of the computer systems we know today, NONE, will be available in 20 years, let alone 50, 100 or 200.

In today's computer world we are dealing primarily with monolithic systems. That is, the hardware, operating system, application, and data are interdependent. IBM or VAX or most any other computer hardware have their own operating systems, their own application, and their own data sets, which can work only within their own world. The perfect archive needs to be accessible not only to all the systems that exist today, but also to all the systems which have yet to be created, all the super fast computers of tomorrow.

Thanks to the Red Book Standard and IS0 9660, CD-ROM is a peripheral device that has already achieved hardware independence. Diskettes, one of the most widely used media in the computer world, will never be able to make this claim. Try putting a MS/DOS diskette into a Macintosh; you can't even get a directory listing, let alone read the files. The 9 track tape might be the only other medium which has achieved the same degree of hardware independence.

Hardware independence is key. An ISO 9660 drive will read any ISO 9660 disk, regardless of the drive manufacturer and machine environment. That is an amazing feat, and again only equalled by the 9 track tape. That is the good news. The bad news is that while hardware independence is a big hurdle, it is only the first hurdle. True system independence demands much more. System independence in today's world means flexible interaction between the following elements:

1. Presentation or Process Software.
2. Retrieval Software.
3. The Information Itself.

This independence can only be achieved by setting standards which define how one element works with another. In essence, this is what the Red Book Standard does at the physical and opto-electronic levels. For logical software transactions, a similar set of standards is needed. Some of these standards have been established, and still more are emerging at this time.

I will limit my discussion to these three standards:

CD-RDx, SGML and TIFF. More should most likely be included, but these will serve the purpose of explaining how such standards work.

Let's start with CD-RDx. This standard was written for the primary purpose of enabling the user to utilize a single interface, his own. To achieve this, CD-RDx uses a client/server approach. The client (the user interface software, or application software) is separated from the server (also known as the retrieval engine). The idea is beautifully simple. If you establish a set of rules by which the client can ask the server for specific "searches" (e.g. a Boolean search), then any client using this protocol can interact with any server using the same protocol. Moreover, CD-RDx is set up in such a manner that the client and the server do not need to be on the same computer, not even on two computers with the same operating system. So now we have system independence between the retrieval engine system and the presentation or application system.

This represents important progress, but system independent data has not yet been achieved. The goal, remember, is to access the data on any disc via any retrieval engine. The solution is to agree on the structuring of information. TIFF and SGML are such standards which could be used to structure information in a universal, non-proprietary way.

Then, a CD-RDx compliant search engine could be developed to access any set of information structured within the guidelines of SGML and TIFF.

There it is, not simple, but resilient. For this to truly work, one also needs to take into account the problem of indexes, though the same reasoning applies, the question of standard index methodology is a truly thorny issue. What we would have is a structure where when one system changes, the information does not need to change and can remain on the same medium.

If the user/client system changes, client software is rewritten to run on the new system, along the guidelines of CD-RDx. If the server system changes, CD-RDx server software is rewritten to retrieve data structured along the SGML and TIFF guidelines, keeping the data unchanged.

This arrangement is being developed today for easy distribution of data. The multitude of clients is due to the multitude of users and potential users of the data distributed, multiplied by the number of titles they each use. In the case of long term archiving, the multitude of systems is even greater because time is a new multiplying factor.

It must also be noted that this solution enables computer user interface and search engine to progress at their own pace, while "stable data sets" stay on one medium. Of course, the characteristics of the medium remain intact. The time will come though, when today's advanced CD-ROM technology will be regarded as a bulky and slow medium. Isn't it a lesser evil though, next to the alternative of losing access to the data content forever?

While today, CD-ROM is the perfect tool to have information on-line and on-site, to be used as an archive, the CD-ROM will be off-line. These archives will be fed into the super computers of tomorrow. After all, the CD-ROM will not always be the medium where the information resides for processing. Often the needed data will be downloaded to fast access, massive storage devices of future computers for processing as needed. This already happens when information is downloaded from a CD-ROM onto a word processor or a spreadsheet.

To recap the longevity factors relating to access systems, we have established that.

Hardware availability in the distant future is all the more likely if the technology is:

1. Widely Accepted- thanks to the support of a product in the consumer market that uses the same technology. Proliferation helping the survival of functioning hardware.

2. The Technology Must Also Be Standardized- which helps the consumer market become even bigger (see 1.) and documents the detailed working of hardware in a precise and widely available manner.

Logical access in the distant future is all the more likely if:

1. Information is system independent.

2. Information is formatted using a non-proprietary standard. Indeed the wide acceptance of a few well chosen standards will foster the development of compatible software and ensure consistent data structuring.

3. The users can have access to this data and then manage it (for display, printing or processing) using the system of their choice through a system independent client/server type of protocol.

More to the point, we have established that CD-ROM is the medium that answers these requirements. The best 9 track tape, while very widely used, does not benefit from the momentum of a consumer market. WORM does not even come into the picture. This is not to say though that WORM, tape and other media, which do not meet some or any of these criteria, will not continue to perform important tasks in our computer rooms.

It is now time to move onto the longevity of the media itself. Of course, given the above conclusion, if one could find a CD-ROM that would last a hundred years or more, we would be home free. As some of you in the audience know, I am with a company that has dedicated large amounts of time and money over the last 5 years to develop such a CD-ROM.

Conclusion

As I see it, CD-ROM, along with fantastic information distribution capabilities, has already more long term archiving features than any other mass computer storage available. Only a few additional steps need to be taken to truly make it one of the top answers to data information archiving for the next 20 years. These steps are the focus of a committee being considered for creation by the Commission on Preservation and Access. All this is very encouraging, and I hope that it can help solve some of your long term archiving needs.

Bibliography:

1

"Taking a Byte out of History: The Archival Preservation of Federal Computer Records", House Report 101-978, 101st Congress, 2nd Session

"GAO Faults NASA for Mismanaging Storage of Valuable U.S. Space Science Data" James R. Asker, Aviation Week & Space Technology, April 2, 1990 (enclosed)

"Lost in Tapes" J. Sniffen, Associated Press, January 2, 1991 (enclosed)

"National Archives Needs Better Record-Keeping Technology, Report Says" Ann M. Mercier, Federal Computer Weekly, November 1990

"'SGML Like'- And We Could Get 'Sort of Married' Too.." William Zoellick, Disc magazine, Premier Issue, Fall 1990. page 53-54. Available from Helgerson Associates. Tel: (703) 237-0682

6“CD-ROM Read-Only Data Exchange Standard, Version 3.0”; December 30, 1990; available from the OPA. Tel: (614) 793-9660

“An Analysis of Compact Discs As a Long term Archiving Solution” Denis Oudard, American Library Association Midwinter Meeting, ALCTS-PLMS' Physical Quality & Treatment Discussion Group, January 12, 1991

“CD-ROM as an Archiving Medium?” Denis Oudard, Working paper, Digipress, February 1991. Available on demand. Tel: (502) 895-0565Presented at the IASSIST 91 Conference held in Edmonton, Alberta, Canada. May 14 - 17, 1991.

[1] Presented at the IASSIST 91 Conference held in Edmonton, Alberta, Canada. May 14 - 17, 1991. For further information on Digipress's Century Disc contack the author at: 2016 Bainbridge Row Drive, Louisville, Kentucky 40207 USA. Tel (502) 895-0565.