CD-ROMs in Astronomy

Kathleen Robertson

Institute for Astronomy Library, Honolulu, Hawaii, USA

Abstract

The CD-ROM has become a major data storage medium in astronomy. The release of the Digitized Sky Survey is the most recent example of this phenomena. I will summarize the large scale jukebox technologies available for making the DSS available over LANs (Local Area Networks).

I will also give an overview of the developments in disciplines such as medicine, chemistry and business, where CD-ROMs have been used to distribute the full text of journals. These factors may give us insights into the future role of CD-ROMs.

When the Guide Star Catalog was the biggest data set out there, an easy solution was to have a dedicated CD drive for each disc. The discs were free and many UNIX workstations came with CD drives for software distribution.

The International Halley Watch set of 24 discs was the first set which began to suggest multiple drives would be necessary for maximum utilization. The set includes several index and correction discs. They must be consulted along with the data discs to insure that the correct values are used. One IfA researcher purchased a second drive immediately after getting the set.

But the 102 disc Digitized Sky Survey (DSS) has made the single CD drive look pathetic. In anticipation of the arrival of the first part of the DSS, I had convinced the IfA budget committee that the library needed a CD-ROM drive for the SPARC station adjacent to the (paper and film) Sky Surveys. Once the first consignment of DSS discs arrived, I was quickly informed by users that the machine was too slow; that when an area of interest crosses several discs, swapping them was tedious; that the SPARC kept "loosing" the CD-ROM driver causing delays and requiring (in the worst cases) rebooting.

With envy, I thought of the rows of terminals providing CD-ROM access at my local university library. If only I could have the CD-ROM contents delivered to the researchers' desks as easily as the Main Library served up ERIC and PsycLit! It would be ideal to deliver the CDs' contents through the network

Preprint submitted to Elsevier Science *5 July 1995*

to individual workstations, many of which were newer and faster than the library's single SPARC 1.

Bibliographic CDs have become ubiquitous in libraries (though somewhat less so in smaller special libraries). Many university libraries have found towers, "stacked" CD-ROM drives with one disc per drive, to be a good solution to providing many users access to multiple disc sets like MEDLINE and COMPENDEX. Towers come in a variety of sizes, from 4 CD-ROM drives to 21 and up. Depending on the type of CPU installed, Virtual's 7 drive tower can support up to 50 users, while their 21 driver tower can handle 150 users. OPTI-NET is near the top in size, supporting access to 255 CD-ROM drives per server, for up to 100 users. (Immediately the plot thickens, if the discs require Microsoft Extensions, simultaneous access drops to a maximum of 24 discs. Many bibliographic CDs do require Microsoft Extensions). These configurations are aimed as giving many users access to a small number of discs. Often these are self-contained systems, requiring hardware and software upgrades to support remote access; most are DOS or WINDOWS based.

Much of libraries' experiences with bibliographic CD-ROMs can't be translated directly to astronomical image CDs. The amount of data moved during a bibliographic session is much smaller than astronomical images. In the astronomy research setting, disc sets can be large; the ratio of discs to users is high, rather than the reverse. UNIX, not DOS, is the prominent operating system.

Like their musical namesakes, CD jukeboxes access many platters (discs) with one or a few players (drivers). CD-OM jukeboxes offer price savings by providing automated access to large number of discs, at prices below the drive-per-disc towers.

CD-ROM jukeboxes can be roughly grouped by size. At the low end are those where one drive accesses six (Pioneer) or seven discs (Mountain Network Solutions, Scotts Valley, CA). Pioneer offers a minichanger with a six disc magazine and two (DRM-602X) or four (DRM-604X) drives. Three magazines can be linked to provide an 18 disc capacity. JVC and Future Echo are expected to introduce new products in the minichanger category this year.

Another group of CD jukeboxes cluster at the 100 disc size. Pinnacle produces the Cascade CD 100 Jukebox (ca. $9,995) and also market a Recordable CD (CD-R) System. NSM currently has a single drive jukebox, CDR 100 XA, that holds up to 100 discs, and has announced a 150 disc jukebox with the option of 2 or 4 drives. Kodak Professional Image Library system includes a 100 disc NSM jukebox and image (scanned page) management software.

Large CD-ROM jukeboxes of 200+ are made by several manufactures including Pioneer, Kubik, NSM and DISC. UMI markets ProQuest, a 240-disc Kubik multi-changer with 4 drives, CDs of journals articles with copyright royalty tracking. But most large systems are sold customized by specialized companies, called system integrators. Firms such as Todd Enterprises, Virtual, and Logicraft market combinations of jukeboxes, towers and communications interfaces that manage the CD-ROMs and link them, via the network, to the workstations. Most large jukebox installations will be a contract between your institution and a system integrator. The many factors involved in the installation make this the easiest route. Atkinson and Yorkley of the Naval Research Laboratory have chronicled their adventures setting up a system from scratch in “Multiplatform CD-ROM Networking” (CD-ROM Professional 1993:73). System integrators can offer much of the same detailed research the authors describe for institutions that don’t have the necessary in-house expertise. Because jukeboxes are so new, many computer professionals have no experience making CD access available through a network.

But, as with any consultant, you must give the system integrators a clear picture of your institution’s needs. When contemplating the purchase of a CD-ROM delivery system in today’s marketplace, one of the most notable things is the diversity of the offerings. No single company or standard dominates; there is no equivalent to the de facto VT 100 standard that has evolved in the computer communication field. Thus, each component in the proposed development must be specified and each offered system must be minutely scrutinized.

When analysing the potential needs for a CD-ROM delivery system, the type and number of CDs must be the starting point. This is the parallel to the given wisdom about selecting computers, what applications do you plan to use. For image CDs, factors to consider include, in what format are the images stored? Are they compressed? What specific software is needed to read them? Is there a statistical, modeling or other routine used in conjunction with the set?

Even if the arrival of the DSS has started the inquiry, it should be remembered in large sets not all discs are the same. There may be index or errata discs that will have much more frequent use than the average member of the set. For those discs, or for heavy use sets like the GSC, non-jukebox access may be needed. A mixed system of jukebox and towers may offer better performance, and/or a large hard disc, to which the data can be cached might give superior performance.

It’s clear that not all the discs you’ll want to make available are going to be image sets. Bibliographic indexes are so plentiful and diverse that some may be added to your collection. It’s becoming increasing common to find backsets of journals distributed in the CD-ROM format. Soon it may become standard practice to receive a quarterly disc, rather than bind paper issues. If long runs of important journals were offered on CD-ROM, rather than microform, I believe I could get funding to enlarge and deepen our holdings. And as acid-based paper begins to deteriorate, scanning to CDs will be an inexpensive solution to the preservation problem. Full text discs can be searched on a word by word basis. Your CD-ROM network will need to support the search engine. But it's unlikely that all will come with the same search interface. So your system must be able to handle the searching programs that come with the products. Microsoft Extensions, required by some bibliographic databases, are an example of the type of add-on that must be anticipated and supported.

Besides the journals stored on disc, a separate, still evolving form is the CD magazine. These products are very new and are currently aimed at the home CD-ROM user. "Time" and "Money" have produced offerings that include video clips, sound recordings, and downloadable tables. There are also a few titles that are published only in CD format, such as NautilusCD, which has a circulation of 12,000, half Windows users and half Mac. As such offerings multiply, some may be appropriate to your collection. I can envision a CD-ROM that combines the text, tables and motion sequences that Astrophysical Journal now publishes separately on paper, disc and videotape.

CD-ROM producers do not always make their products "network aware." Some are designed for a single workstation environment. When adding new offerings to an existing network, demand a 30-day evaluation copy to test network compatibility. In the planning stage all you can do is try to identify and itemize the various applications that need to be supported now.

Future developments will include, but not be limited to, erasable CD and higher density CDs. An agreement on a format for erasable CD-ROMs has been reached by 10 leading computer and media companies. The erasable CDs will become available in 1996, making it possible for companies to store and revise large amounts of data. The new format is to be "fully compatible" with existing CDs. Drives will require only "minor modifications" (Wall Street Journal May 2 1995, B6). The implications for CD-ROM networks and jukeboxes is unclear. Even now much of the market for large jukeboxes is driven by the insurance companies, financial institutions, etc. that are selecting CD-ROM as the medium for their archived records. No industry-wide standard for higher storage densities has yet been established, though several have been proposed.

Because of the pace of developments, it might be a mistake to select a very large jukebox and wait for it to be filled solely by the pace of published CDs. If your institute has CD-R capacities, you'll be able to plan for some in-house resource growth. But, I have only about 140 CDs in-hand. I received a quote on a DISC (JBD300-1, 4 Toshiba drives) jukebox that has a 333 disc capacity. If I selected that, I'd be assuming that the no major change in disc format will come quickly. It's safer to select a jukebox in the 200-240 range. Then add another jukebox to handle new formats, when the need arises.

A major consideration in CD-ROM networking is the existing Local Area Network (LAN). Appleshare, Banyan Vines, LANtastic, Microsoft, NFS, UNM, Windows for Workgroups are among the networks currently in use, but Novell is the most numerous.

Consideration should be also be given to possible LON (LAN Outer Net), connections including branches, home access, mobile users, etc, as well as WAN (Wide Area Network) support for remote dial-in access. Questions such as how many concurrent dial-in users can access simultaneously, is remote downloading supported, and how are "hung" remote sessions cleared, should be considered.

Speed of data transport is another issue. It's influenced by both the jukebox mechanism and by the LAN bandwidth. Jukebox manufactures usually give an average access speed estimate. Kubik advertises "12 second average disc load time" for its CDR-240M model. This should be verified by tests, both with an open driver and with another disc that must be swapped out of the driver. The slowest segment of the LAN will determine the speed of delivery. Data may move at 100 Mbps (mega bits per second) over the FDDI backbone, but will slow to the 10 Mbps of your local Ethernet link.

Today, 80% of CD-ROMs are DOS based and an additional 15% use Windows. Mac represents only 3% of the offerings, UNIX only 1%. Because of the high portion of CD-ROMs that are DOS based, consideration must be given to access for non-DOS machines. The system integrator's software must support translation between the DOS-based CD and the user's screen. Does it also support the ability to map the users' keyboards to the DOS application keys. (Think of Silver Platter's use of the F4 key to initiate printing).

For UNIX based LANs, several options are available. TCP/LAN Gateway software can be loaded on the UNIX host. Users will telnet using X11 protocol; PC and Macs will run under XProtocol. The NFS option consists of a high-end UNIX host such as a Sun Server (or the server running "SCSI Express for UNIX" and NFS, or "Netware NFS" (NLM) and "SCSI Express" (NLM)). DOS and Windows workstations run PCNFS or a product such as Chameleon by NetManager. Mac and UNIX workstation run NFS, then run DOS emulation. "Young Minds" supply drivers that permit each disc to appear as a sub-directory off a common NFS drive mount. From the TCP/IP backbone, UNIX and Macs run NFS with DOS emulation. To connect LANs in different locations, "T1" or "T3" leased lines are used, linked by routers.

Mention of TCP/IP raises the question of Client/Server connections in a CD-ROM jukebox system. The client/server approach reduces backbone and WAN traffic. But if the link is a 56 Kb/s (kilo bits per second) WAN link, speed will degrade to unacceptable levels. On faster lines, the UNIX host provides the DOS Merge session and transport software (SCSI Express) links, and the user telnets to the UNIX host. Another approach is to use a concentrator (black box) and specialized software ("EveryWhere Access") to support the variety of terminal types. The user needs only telnet in.

In addition to the communication function, system integrator software should support systems management functions. Does the package monitor CD-ROM system usage, does it meter access to licenced CD for contract compliance? This function is built in to OPTI-NET, for examples. Others add 3rd party applications to do this task. Does the package manage time-out functions to end abandoned sessions? Does it support a front-end menu for easy resource selection by users (standard with DOS, not a given with UNIX). Does it support remote downloading? remote printing?

Once the system is installed, fine tuning may increase response time. Where should the heavy use discs be placed for optimum speed? Which discs should be mounted in the tower? Which cached to a hard disc?

As with any equipment purchase, checking the vendor's track record and talking with the vendor's clients can help in RFQ preparation and system selection.

Write up your contract so that the burden of providing performance is on the system integrator. Don't be sold a bigger system than you'll use in the short to medium term. Look for hardware options that allow flexibility for future developments. And, let me know how your installation goes and how the system performs, so I can benefit from your experience.

One last point, as CD-R capacities become common, the possibility increases of CD-ROMs becoming an area of gray literature. When everyone can master their own CDs, basic information may not be attached to the physical disc. Urge your researchers to give complete publication data in bibliographies and to label discs for accurate citation.