An Introduction to Optical Media Preservation

by alex duryee
digital & metadata preservation specialist
avpreserve

AVPreserve | Media Archiving & Data Management Consultants
350 7th Avenue, Suite 1605 | New York, New York | 10001
917.475.9630 | www.avpreserve.com | info@avpreserve.com

Introduction

The role of optical media in archives has shifted in the past decade from preservation medium to at-risk format. While longevity models estimated a lifespan of 25-200 years for recordable media[1], recent testing has found this range to be optimistic by orders of magnitude. One collection of data CD-Rs from the 1990s yielded a 92% failure rate after approximately twenty years of storage[2]. Given how recently archives have approached optical media as an object of preservation, there is little literature and research regarding the format. While a comprehensive account of all migration and preservation issues of optically stored data is impossible, there is a need for a broad overview of optical media, which this document hopes to provide.

This document will cover five of the most common data storage standards for optical media. In doing so, it hopes to inform institutions working with such media as to the nature and challenge of optical media with regards to preservation. The author also hopes that, by providing a general overview of optical media preservation, more advanced conversations and explorations of the medium can take place.

Note that this article deliberately limits its scope to the most common formats and uses of optical media. The total forms and functions of optical media are many and varied – from analog video to hidden data to graphics stored in control codes – and are beyond the scope of this document. Archivists dealing with such discs are recommended to explore more specialized resources, such as technical communities and standards documentation. This document, while offering recommendations, is also not a prescriptive cookbook of workflows and tools for dealing with optical media, as the necessary research for such recommendations has yet to be performed.

Logical Structure

It is crucial to understand the logical layout of optical media before attempting any preservation activities. As the first optical media standard was IEC 60908 (1982) for audio storage and playback, future standards reflected a media-centric approach in how they structure the disc. Thus, despite standards such as ISO/IEC 10149 (available as ECMA-130[3]) providing for filesystem-based storage, the language and structure of Red Book persists[4]. Hence, while it seems strange to discuss file-based data in terms of logical tracks, this is precisely how data is stored on CD-ROM.

A simplified model of a compact disc's logical structure is as such: a series of sessions, each containing a series of tracks. A track is, as its name implies, designed to be one discrete track of audio. For non-audio data (e.g. CD-ROM), any number of filesystems may be contained within a single track. These tracks are arranged in a linear series and bounded by a lead-in (which contains the table of contents – locational and descriptive metadata – for the following tracks) and a lead-out. This collection comprises one session. Early CDs were designed as if only one session would be on a disc; this was expanded in 1990 to provide for multiple sessions.

[1] Range Commanders Council, Optical Systems Group. Multimedia Archiving: Videotape, Compact Disc (CD), Digital Versatile Disc (DVD), and Blu-Ray Disc (BD) Media [Internet]. White Sands Missile Range, New Mexico: Range Commanders Council [updated 2010 February; cited 2014 February 14]. Available from http://www.wsmr.army.mil/RCCsite/Documents/462-10_Multimedia%20Archiving%20-%20CD,%20DVD,%20and%20Blu-ray/462-10_Multimedia%20Archiving%20-%20CD,%20DVD,%20and%20Blu-ray.pdf

[2] Wilsey L, Skirvin R, Chan P, Edwards G. Capturing and Processing Born-Digital Files in the STOP AIDS Project Records: A Case Study. Journal of Western Archives [Internet]. [cited 2014 February 14]; 4:1. Available from http://digitalcommons.usu.edu/cgi/viewcontent.cgi?article=1026&context=westernarchives

[3] Ecma International. Data interchange on read-only 120 mm optical data disks (CD-ROM) [Internet]. Second edition. Geneva, Switzerland: Ecma International. [updated 1996 June; cited 2014 February 14]. Available from https://web.archive.org/web/20130116084229/http://ecma-international.org/publications/files/ECMA-ST/Ecma-130.pdf

[4] The Red Book is the colloquial name for IEC 60908, due to the color of its cover. The various standards that define optical media are collectively known as the Rainbow Books, as each has a differently colored cover (e.g. ISO/IEC10149 being the Yellow Book).

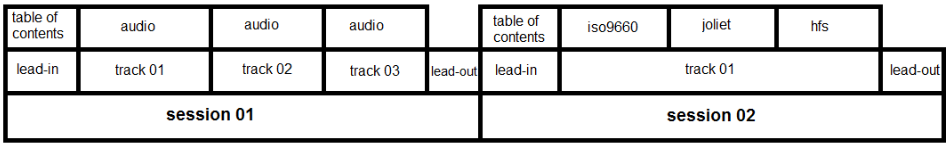


Figure 1. A diagram of the logical structure of a multi-session CD.

Due to the flexibility of how data can be arranged on sessions and tracks, the following example configurations are possible:

- One session, many audio tracks (typical audio CD)
- One session, one track, many filesystems (typical data CD)
- Two sessions, one with many audio tracks, one with one track and many filesystems (Enhanced CD)
- One session, one data track, many audio tracks (video game)

It merits noting that, on CD-R and CD-RW, a session can be open or closed. A session can have new tracks written to it so long as it is marked as open; by closing it, it prevents further data from being added and creates the final table of contents. Methods of handling discs with open sessions are documented by the digital forensic community, due to their importance in evidence collection[5].

For purposes of this article, the structure of DVD is sufficiently similar to CD-ROM that it does not merit in-depth analysis.

ISO9660

ISO9660 is the most common format for carrying data on CD-ROMs, often appearing as the baseline filesystem on cross-platform discs. It is also common for it to serve as the foundational filesystem for more sophisticated systems, such as Joliet and HFS – these often (but not always) serve as a layer on top of ISO9660 data to provide additional functionality. Its limitations (such as 8.3-style non-Unicode filenames) make it compatible with all common operating systems, thus allowing for a fallback in case more advanced filesystems are not supported. As such, it is extremely common to find compact discs with one ISO9660 filesystem alongside a mix of HFS and Joliet – such a disc would provide advanced features for Windows and MacOS, but the data would be accessible on virtually any operating system.

From the perspective of preservation, ISO9660 can be treated similarly to filesystems on magnetic media. As each sector must return data consistently on each read (consider the consequences of a dropped sector when executing compiled code!), and each sector must be easily found on the disc, these discs use the Mode 1 sector structure. In contrast to CD-DA's use of every byte in a sector for data, Mode 1 dedicates 2048 bytes to data, 16 bytes to sector sync and identification, and 288 to error detection and correction. These non-data bytes reduce the amount of user data on the disc but crucially allow for it to behave as a filesystem by providing for rapid accurate sector seeking and consistent data reading.

[5] More information on the theory and method of retrieving data from discs with open sessions can be found in Crowley and Kleiman's CD and DVD Forensics

As such, for purposes of migration, long-term preservation, and access, an ISO9660-based disc can be treated similarly to a magnetic disk. While there is no need for the specialized hardware that magnetic media requires (such as write blockers), ISO9660 tracks can be imaged at the byte level[6] via the standard variety of data transfer tools (dd, guymager, FTK, IsoBuster, etc.)[7]. The subsequent image can then be mounted as any other filesystem for archival analysis and user access. Various access models and post-migration workflows, such as remote image mounting and automated filesystem analysis, have previously been explored similarly and parallel to research in magnetic media[8].

Joliet / HFS

While ISO9660 is the most common data format for CD-ROM, it is rarely seen by itself. Due to the restrictions on file structures, there was a demand for some mechanism to expand the capabilities for CD-ROM filesystems. To address this, Microsoft established the Joliet specification as an extension to ISO9660 in 1995. Joliet's primary improvements over ISO9660 are long filenames, Unicode filenames, and deeper directory trees. As Joliet does not manage the data on the disc – it only provides for enhanced filesystem metadata – it is typically packaged as a metadata layer next to an ISO9660 filesystem. An exploration of a disc's logical structure will reveal that an ISO9660 file and its Joliet counterpart both point to the same sector of the disc – while the filenames differ, one copy of the data is used.

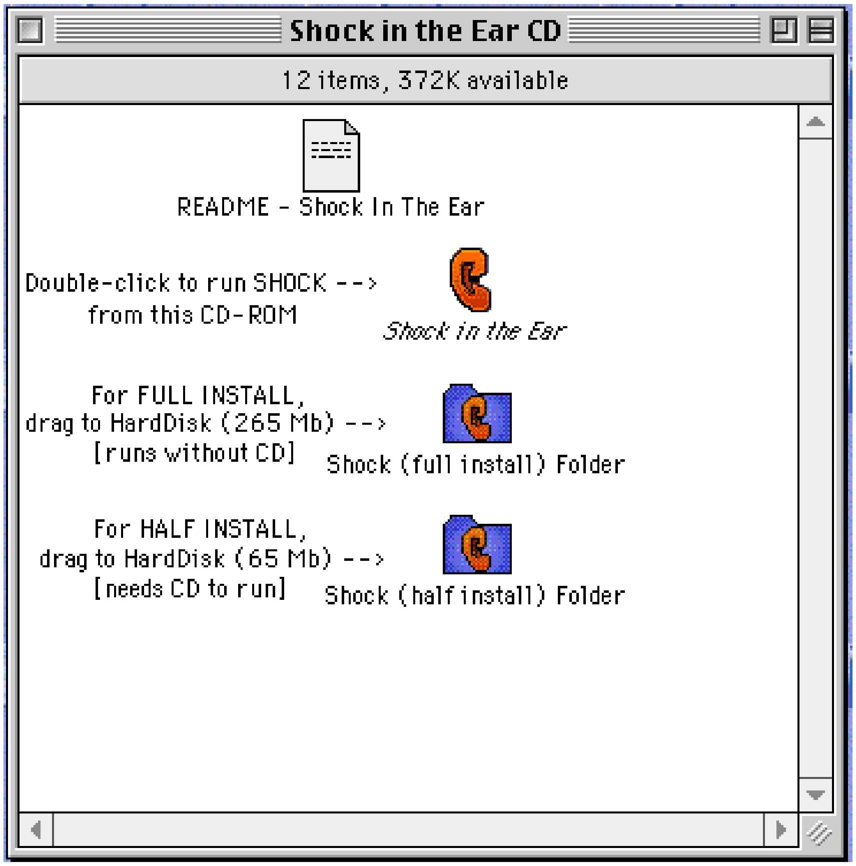
The Hierarchical File System (HFS) was employed to allow for Macintosh-specific file behaviors on CD-ROMs and to work around the limitations of ISO9660. HFS, while specific to Apple machines, is a more powerful filesystem than ISO9660 – notably, it allows for longer filenames and metadata necessary to integrate more naturally with the operating system[9]. As such, it was heavily used to provide compatibility with MacOS; for example, a document without the necessary HFS-specific metadata will lack the format/software signifiers to open correctly. Thus, even in environments where the HFS filesystem is no more than metadata pointing at underlying ISO9660 data, the HFS filesystem is critical in retaining the compatibility of the disc. Typically, HFS filesystems will contain data unique to the disc on MacOS, such as compiled bytecode and documentation specific to the operating system, which should be migrated as part of any archival workflow.

[6] Note that the term "bit level" is meaningless with regard to optical media. Instead of using eight physical bits per byte, all compact discs use eight-to-fourteen modulation (EFM): the 256 combinations of fourteen bits most favorable to accurate reading are mapped to corresponding eight-bit bytes, and the fourteen bit words are written to disc. As such, a true "bit level" rendition of a compact disc would require an analysis of the physical artifact. For more information, see the ECMA-130 standard linked above.

[7] Note that the colloquialism "ISO image" is frequently misleading. Despite implying that a disc image file ending in .iso captures only ISO9660 data, it is instead used to describe a general disc image. As such, unless the provenance of a .iso file is documented, it cannot be used when determining the scope and content of migration.

[8] Woods, Kam A. (2010). Preserving Long-Term Access to United States Government Documents in Legacy Digital Formats [dissertation]. Bloomington (IN): Indiana University. Available from https://web.archive.org/web/20140217161345/http://www.digpres.com/publications/Kam-Woods-Dissertation-Pre.pdf

[9] For example, the creator code and type code allow for MacOS to open a file automatically using the intended software.



Shock in the Ear CD

12 items, 372K available

README - Shock In The Ear

Double-click to run SHOCK -->
from this CD-ROM

Shock in the Ear

For FULL INSTALL,
drag to HardDisk (265 Mb) -->
[runs without CD]

Shock (full install) Folder

For HALF INSTALL,
drag to HardDisk (65 Mb) -->
[needs CD to run]

Shock (half install) Folder

Figure 2. A disc containing an HFS filesystem, viewed in an emulator. Note the floating text in the shell – this was accomplished by creating empty files with a blank icon and placing them within the Finder window.

UDF

Since UDF was designed as a universal standard for data storage, it has few surprises for preservation. As opposed to a compact disc, a DVD will almost certainly (barring unusual circumstances) behave in a standard fashion across any number of devices. Despite the apparent variations between DVD-Video, DVD-Audio, and data DVDs, the underlying disc format is identical. While there are a variety of quirks between versions and implementations of UDF and highly specific cases where its structural features are critical to a successful migration, these are beyond the scope of day-to-day preservation.

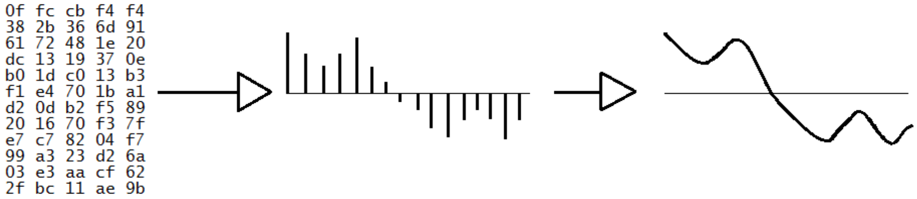
Note that, due to varying hardware support for UDF versions, it is often paired with an ISO9660 filesystem. Due to ISO9660's near-universal support, it can be used as a “bridge” to allow older hardware to read the disc. As such, it is up to the individual archive to decide if one or both of the filesystems should be preserved.

One important subclass of DVDs is DVD-Video, which is the predominant format for consumer video. Using a screen essentialist approach, DVD-Video is seemingly equivalent to audiovisual tape formats: the user inserts the disc into a player, which provides an interactive menu and a collection of linear audiovisual streams. However, this masks the underlying structure of the disc. In actuality, the DVD contains a UDF filesystem with a standardized directory and file structure (which a DVD player recognizes and parses in a seemingly single stream). Put generally, the filesystem contains a VIDEO_TS directory, which contains the MPEG streams

(VOB) and playback metadata (IFO/BUP)[10]. DVD-Video can therefore be treated similarly to any data DVD for purposes of migration. The major issues that archives will face with DVD-Video are overcoming the Content Scramble System (a copy protection scheme that prevents disc migration and hence preservation) and providing access to the raw audiovisual files post-migration.

CD-DA

The data structure of a CD-DA is more akin to tape than a traditional data disk. Instead of dividing the storage area into discrete files, CD-DA data is written as a linear pulse-code modulation (PCM) stream, divided into separate tracks. As the data is read by the disc drive (at 44,100 16-bit samples per second, chosen specifically to be above the necessary rate for perfect human reconstruction), the playback device interprets the PCM stream and generates the corresponding waveform. In this regard, CD-DA is more closely allied to tape-based formats than it is to traditional magnetic disk, as it represents not a structured filesystem but a linear stream of media.



```
0f fc cb f4 f4
38 2b 36 6d 91
61 72 48 1e 20
dc 13 19 37 0e
b0 1d c0 13 b3
f1 e4 70 1b a1
d2 0d b2 f5 89
20 16 70 f3 7f
e7 c7 82 04 f7
99 a3 23 d2 6a
03 e3 aa cf 62
2f bc 11 ae 9b
```

Figure 3. The pulse-code modulation stream to digital signal to analog signal chain

As CD-DA is fundamentally different from data formats like ISO9660, there are a number of factors to consider. Since there are no ‘files’ on a CD-DA session, but streams of raw PCM data, the file listing provided by an operating system’s shell does not reflect the contents of the disc. Windows, for example, will display .cda files that will play via media software. These files, however, contain no data – they are merely pointers to locations on the CD where their corresponding tracks begin. The data itself is something that cannot be parsed by the shell and thus requires special software to migrate.

CD-DA was also designed to maximize space at the cost of accuracy; thus, it sacrifices a third level of error correction in exchange for more data per sector[11]. The advantage of this approach was a marked increase in capacity and a level of error tolerance, as misread sectors will not interrupt smooth playback in all but the most egregious cases. Thus, errors introduced during the read process can go undetected by the drive. Typically, read errors will manifest as traditional sonic problems, such as clicks and pops.

As a result, reading an audio track in a single pass will provide unreliable results, with consumer hardware being roughly 95% accurate at the track level[12]. This is unacceptable for preservation – consider the consequences of introducing errors to five to ten percent of tapes and records during transfer! Various techniques have been developed to account for CD-DA’s inherent

[10] The DVD Video specification also defines an AUDIO_TS directory. This is only used on the exotic DVD-Audio format and will likely not be encountered in most archival scenarios. For the purposes of DVD-Video, it will most always be empty.

[11] The specifics of error correction of CD-DA data is complex enough to fall beyond the scope of this document. More information can be found at: https://web.archive.org/web/19970616191108/http://www.ee.washington.edu/conselec/CE/kuhn/cdmulti/95x7/iec908.htm

[12] CD/DVD Drive Accuracy List [Internet]. [updated 2013 May 14]. [Cited 2014 February 14]. Available from https://web.archive.org/web/20131203200732/http://forum.dbpoweramp.com/showthread.php?30430-CD-DVD-Drive-Accuracy-List-2013

lossiness. Software designed for reliable CD-DA extraction combine a variety of methods to ensure consistent reads across a given disc. For example, an extraction tool may read each sector of the disc multiple times in order to find the correct value, or it may compare the data against online databases containing other bitstreams of the same disc.

While these methods are powerful in ensuring consistently good reads of CD-DA, they are specific to this data structure, and thus they are not portable to ISO9660 or other CD formats. This presents issues during the migration of so-called Enhanced CDs (ECD, also known as CD Extra) and mixed mode CDs (common with video games), which contain both audio and filesystem data. As the specific workflow for such discs is reliant on the structure of a particular disc, it is beyond the scope of this document. As it is unlikely that these will be found outside of specific subject collections, archives handling such media would do well to conduct research and contact specialists with regard to migration.

Applications in Preservation Workflows

The most important analysis an archivist can perform on a disc occurs before it ever enters a workstation. Recordable optical discs are frequently used by individuals as simple portable file carriers – for example, a donor may burn a DVD with material to transport as part of their collection, or a disc may be created to share documents around an office. In these cases, the disc serves as a simple carrier, analogous to an envelope or package – in other words, something that never merits preservation. As such, the archivist may decide to copy the files over via their shell (Finder, Windows Explorer, et cetera) instead of imaging a disc, as there is no value in preserving the low-level structure. Dependent on the nature of the discs in a given collection, this may be the preferred method of migration.

The physical aspects of optical media, while typically not important to the data on the disc, do merit discussion with regard to migration processes. The read speed of a disc should be set as low as possible (via tools such as CD-ROM Tool SPTI or hdparm), as lower speeds can provide more accurate results. The quality of drive can also make a dramatic difference – empirical data has found high-quality consumer drives to be approximately five percent more accurate at the track level than poor ones[13] – and thus should be taken into account when designing an archival workstation. Traditionally, Plextor drives have served various preservation-minded communities well and therefore often recommended for optical migration[14].

Given that an optical disc can have any number of filesystem and sector structures – some of which are invisible via an operating system's shell – it is necessary to use dedicated tools for analysis. By reading a disc's table of contents directly (instead of relying on what the operating system recognizes), it becomes possible to view a complete list of tracks on a disc (and their corresponding contents in case of data tracks). A handful of tools exist for exploring the structure and contents of a disc, such as CD/DVD Inspector and IsoBuster (both Windows-only). These will allow for an archivist to understand the broader structure of a disc (sessions/tracks, filesystems, files, etc.), as well as catching any discs containing CD-DA before migration.

[13] *Ibid*.

[14] Note that Plextor ceased production of its own hardware in the mid-2000s; any drive with the Plextor name made after that point was manufactured by a different company.

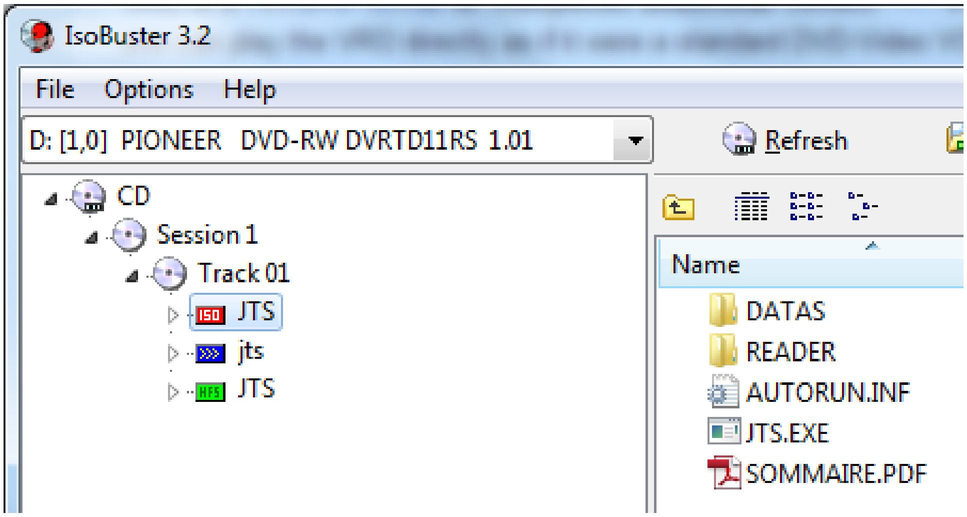


Figure 4. IsoBuster with a disc containing an ISO9660 filesystem, extended by Joliet and HFS layers

Due to CD-DA's lack of error correction in comparison to other CD/DVD formats, it requires a special workflow to migrate properly from CD. A linear read of a CD-DA track will not give an accurate transfer of data, due to reasons outlined above; additionally, audio-specific issues (such as placement of silence between tracks) are issues not addressed in more general software. To solve these problems, specialized software was developed to address problems in CD-DA extraction. The author recommends tools such as Exact Audio Copy (Windows), dBpoweramp (Windows) and cdparanoia (Linux/OSX/Windows), which are specifically designed to overcome migration barriers via methods to account for silent read errors, pregap detection, et al. The combination of empirical data and technical analysis of these tools make them the gold standard for audio CD migration.

The target formats for optical migration are dependent on a variety of factors. A complete image containing every byte on a disc will not necessarily be the optimal format for long-term preservation, as it will include error correction and sync data. This will increase the size of an image by 305 bytes per sector (96MB across the entire disc), provide minimal utility, and create potential access issues. However, by using a binary image (typically a .bin file) and a cue sheet (.cue file, which provides the metadata necessary to divide the binary stream into sessions/tracks and interpret it), a disc can be replicated perfectly. In certain cases, particularly that of complex discs, this may be necessary for creation preservation masters. An alternative when handling single-track discs is a single track image, which will capture the user data within the track and provide a mountable and usable image. General workflows for the creation of preservation masters can be found in the Appendix.

For CD-DA, the target migration format is typically 16-bit/44.1kHz WAVE. As the raw PCM data is equivalent to a headerless WAVE file, the header can be appended with no change to the audio data. This file can then be manipulated to an archive's specifications for transcoding, Broadcast WAVE metadata, et cetera. If the structure of the disc is important, the bin/cue format is an alternative to discrete WAVE files that retains the timing and layout of the disc[15]. A suggested workflow using Exact Audio Copy can be found in the Appendix.

As DVD-Video is structured no differently than any other data DVD, a binary image of the disc is

[15] The canonical example of bin/cue being necessary for CD-DA is music designed for dance, which often uses no gaps between tracks. As bin/cue allows for direct manipulation of gaps, this allows for the session to be played back as intended.

sufficient for preservation. Given that there is no real need for transforming the MPEG streams within the VOB files for preservation, the image can be stored as-is and mounted upon request for access. However, due to the variety of needs and methods for preserving, describing, and accessing digital video, it is impossible to suggest a single target format – for example, an asset management system may require a specific codec for playback. While the VOB files are stable and generally playable, there may be a need to transform the files for broader/easier access (e.g. concatenating split files into one object). An archive will need to set policy for handling copy protection, transcoding, storing, and providing access to the media.

Future Research

The current state of preservation research with regard to optical media is rather lacking, as it limits its scope to the physical artifact. While relevant in the short term, the inevitability of disc rot, coupled with the decrease in the manufacture of quality drives, places an urgent imperative on migration research and practice. As there has been little research performed in allied fields such as law enforcement and computer science, there is a greater need for the establishment of practices and knowledge relating to optical media by archivists.

Areas of future research may include:

- Documentation of best practices with regard to known and unknown types of optical carriers
- Establishment of techniques and workflows for damaged media
- Scaling optical media migration and balancing best practices with efficiency
- Outlining of significant properties of classes of optical media
- Metadata standards for describing optical media and data stored as such
- Exploration of methods for preserving more exotic optical formats, such as CD+G and LaserDisc-based formats

The author anticipates that, as the archival scope begins to encompass the era of optical storage, a greater need for established workflows and advanced research will arise. The use of optical media as objects of archival preservation has been limited to specific projects (namely government documents and digital art) and external communities (particularly music and video game collectors). By applying and adapting existing knowledge to generalizing optical preservation, archivists can prepare for the next generation of digital preservation challenges.

Appendix

CD-ROM Suggested Workflow

- Analyze disc with ISOBuster and determine workflow (with a particular focus on CD-DA, number of tracks/sessions, and initial errors)
- If the disc is CD-ROM only, analyze for structure (ISO-based, HFS, hybrid, etc) and describe as per metadata standard
- Extract CD <Image> -> User Data, or if non-data bytes are necessary (for byte-level alignment), RAW

This will capture all CD-ROM sessions/tracks on a disc in a binary image. If User Data was selected, this can be mounted as a disc within one's operating system; if RAW was selected, one may need to extract the user data before mounting the image. RAW images can be translated into mountable user data via the bchunk tool on Linux/OSX.

CD-DA Suggested Workflow

Due to the complexity – and necessity – of properly configuring Exact Audio Copy (EAC), it is beyond the scope of a general document. A guide to the varying software and drive options can be found at the Hydrogen Audio Knowledge Base[16]. Note that, while EAC's software options can be generalized for preservation, every drive is unique and thus requires individual configuration. It is also prudent to align one's drive with the AccurateRip database in order to compare individual drive accuracy against others of the same model.

- Analyze disc with ISOBuster and determine workflow
- If the disc is entirely CD-DA tracks, open EAC and read the disc into it
- Detect pre-gaps and check for unusual deviations (gaps are typically approximately 2 seconds each)
- Test & Copy Selected Tracks -> Uncompressed

This will produce WAVE files as per specifications set in EAC's preferences. Due to EAC's emphasis on reliably reading CD-DA (via methods such as comparing multiple reads against each other), this will be a very accurate migration of the data on the disc. If there is a need to maintain the structure of the disc via a cue sheet, this can be created within EAC using the current gap settings.

Glossary

CD: Compact Disc. A format designed to hold audio data (and later expanded to general data) on a 12cm plastic disc, using a laser to read a series of pits and lands as binary data.

CD-DA: Compact Disc – Digital Audio. The standard used to define the logical, physical, and data structures of audio discs.

CD-R: Compact Disc – Recordable. A WORM (write once – read many) format allowing for compact discs to be created using consumer hardware.

CD-ROM: Compact Disc – Read Only Memory. The standard used to define the logical, physical, and data structures of data discs.

CD-RW: Compact Disc – Rewritable. A format, similar to CD-R, that allows for data on the disc to be written over.

DVD: Digital Versatile Disc. A format designed to store cinematic-length motion pictures on a 12cm disc. This led to the creation of a disc that offered 6-10 times the capacity of CD-ROM in the same form factor.

DVD-Video: The standard filesystem and file formats for distributing motion pictures on DVD.

Enhanced CD: A format for storing CD-DA and CD-ROM data on a single disc. This was originally impossible, but was permitted with the Blue Book standard in 1995.

HFS: Hierarchical File System. The standard filesystem used on MacOS, which was used on CD-ROM to provide compatability and enhanced features.

ISO9660 ("ISO"): The standard used to define the filesystem and file structure in place on

[16] EAC Options [Internet]. [updated 2013 August 3]. [Cited 2014 February 14]. Available from http://wiki.hydrogenaudio.org/index.php?title=EAC_Options

compact discs. Despite being highly limited, it is near-ubiquitous on CD-ROM.

Joliet: An extension to ISO9660, designed by Microsoft, that provides a metadata layer to overcome the limitations of ISO9660.

Mode 1: A substandard of the CD-ROM data structure. Mode 1 defines 2048 bytes of user data per sector, with the rest of the sector used for error correction and sync bytes. Mode 1 is much more commonly used than Mode 2, which was designed for use in more exotic formats such as CD-i and Video CD.

PCM: Pulse-code modulation. A method for storing analog audio data as a series of binary values.

UDF: Universal Disk Format. A standard designed to supersede ISO9660 by lifting some of its restrictions. UDF set out with the goal of replacing the myriad CD data formats with a single one.

This paper was original published in the *Code4Lib Journal*, Issue 24, 2014-04-16, ISSN 1940-5758. http://journal.code4lib.org/articles/9581