Disk Image Format Primer

Last Updated: 2019-10-25

Purpose of this Primer

This primer aims to give a broad overview on the practical uses of disk images as they pertain to archival practices and workflows. It is by no means comprehensive, and we recommend seeking out more information on digital preservation techniques, software, and hardware if this is a topic that interests you or is useful/might be useful in your line of work.

What is a Disk Image?

According to Wikipedia, a disk image is

> "a computer file containing the contents and structure of a disk volume or of an entire data storage device, such as a hard disk drive, tape drive, floppy disk, optical disc, or USB flash drive. A disk image is usually made by creating a sector-by-sector copy of the source medium, thereby perfectly replicating the structure and contents of a storage device independent of the file system. Depending on the disk image format, a disk image may span one or more computer files."

Disk images, when mounted by a host computer, simulate the physical media they are cloned from.The images are bit by bit recreations of the digital media they are cloned from. They are typically mountable within a operating system (such as Windows 10, Max OS, or Linux) as read only, and files can be extracted from them using forensics tools and software, such as Forensic Toolkit, or the imaging tools found in the Bitcurator environment. Proprietary image formats typically require special tools or software to extract the files and information from the disk images.

A note on terminology: Disk vs Disc:
A "Disk" refers to a form of magnetic storage that is the basis for floppy disks as well as hard disk drives (HDD). A "Disc" refers to optical media, such as a CD-ROM, DVD or Blu-ray disc. In both cases, a disk image refers to the digital copy of the contents of the original physical media.

Why are Disk Images useful?

For archivists and record managers, a disk image allows us to interact with and examine the contents of a disk in a way in line with preservation goals and guidelines. It retains the data structure and contents of the original disk, copied at the bit level, including any file metadata as it was at the time the disk was imaged. This allows us to maintain provenance and original order. It also provides an easily verifiable point from which access copies (in the form of either copied disk images or extracted files) can be derived and reproduced should the integrity of the current access copies ever be compromised. Unrelated to archival practice, they also prove useful in forensic investigations, and much of the technology developed to create disk images originates from that field of work.

How are Disk Images Created?

Depending on the original disk which is being imaged, there are different methods and approaches to imaging disks. One of the most important steps to take when creating images for archival or forensic purposes is to ensure that our host computer does not create or alter files on the disk being imaged. This can be accomplished via a variety of methods.

For optical media (CDs, DVDs, Blu-ray disc) accidental writing is unlikely. The lasers used to read discs are significantly less powerful than those used to write. Imaging can be completed by a standard optical media drive. Although it should be noted that consumer grade drives vary widely in quality, and if one fails to successfully read a disc, another might be successful.

For floppy discs (3.5,' 5.25',8'), an external floppy controller with write blocking capabilities is used in addition to a legacy floppy drive to create a bit by bit image of the original disk. These external controllers (such as the Kryoflux or FC5025) are delicate and volatile, and care needs to be taken while using them. Additionally, floppy media typically included physical write blocking methods that vary by specification. These can be employed for additional security.

Contemporary mass storage drives (USB, HDD, SSD) can be imaged using open source software (such as Guymager, which is included in the BitCurator environment), as well as using forensic software (such as Forensic Toolkit or EnCase). However, it is important to employ physical and, if possible, software based write blocking methods to ensure that the host OS does not alter or

create miscellaneous files on the cloned drive during the imaging process.

Media Type	Optical (CD, DVD, ETC)	Floppy Disk (3.5', 5.25', 8')	HDD, SSD, USB (flash)
Imaging Software	ISObuster, ExactAudioCopy, Guymager	Kryoflux DTC, FC5025 software	Guymager, FTK, EnCase, etc.
Imaging Tools	Optical Drive	Floppy controller + legacy floppy drive	Physical write blocker, USB connection
Write blocking	N/A	Software/ Write block on floppy controller, media write blocking	Forensic bridge write blockers (such as Tableau), software write blocking
Image Format(s)	.ISO, .WAV and BIN/CUE (For audio)	RAW bitstream formats (DD, IMG), Forensic Formats	RAW bitstream formats (DD, IMG), Forensic Formats

A Note on Bitcurator:
Bitcurator is a set of scripts and tools distributed within a Ubuntu (linux) distribution. It includes forensic reporting and imaging tools designed to assist in the development of archival workflows and processes. Many of these can be used as the software component of generating a Disk Image. It can be run as a discrete system, or as a virtual system using software such as VirtualBox. Many of the component applications have Windows or MacOS distributions as well. More information can be found on the Bitcurator project page and github.

Types of Disk Image formats

When we create disk images, they can be saved in a variety of formats, depending on both the nature of the original media, as well as the intended purpose of the image. The types of formats that exist range from open source specifications to proprietary formats designed and managed by technology companies.

Many of the disk image formats that exist today were originally created for forensic criminal investigations. As such, these formats have a fair amount of documentation and adoption by the communities they were originally intended for. For digital preservation purposes, these formats can have various benefits and detriments, depending on the needs of an archive, such as including additional descriptive metadata about the creation of the image.

The other major category of Disk Image formats are RAW formats. These formats create a bit by bit replica of the original disk without any specifications, compressing or conversion. As long as the underlying file system is known and understood, obtaining access files is straightforward. These images are open source and well understood and documented. The major potential for complex issues lies with unsupported and legacy file systems, which are not a trait of the RAW image itself.

A Note on File Systems:
While a detailed discussion of file systems is beyond the scope of this primer, it seems pertinent to note what a file system actually is and does for a disk. File systems determine the logical organization of information and files on a disk. Since data is not always stored sequentially, having technology that is compatible with the underlying file system is important when creating a RAW disk image.

A Note on Logical Disk Images:
Logical disk images only creates a copy of files and data that is detailed by a disk's underlying file system. When computers 'erase' files, they typically do not overwrite the information immediately. Instead they unallocate the space where those files were stored, which allows them to be treated as free space by the operating system. A logical disk image would not include these unallocated fragments, which are something that archivists might wish to preserve. One benefit of a logical disk image is that it does not include all of the free space on a disk.

Pros and cons of selected Disk Image formats

	AFF4	EWF-E01 (EWCF-ASR02)	RAW (DD, IMG)	ISO
Open vs Proprietary	Open	Proprietary file owned by Guidance	Open/Not a specification	Open/Not a Specification

			Software, Inc.	
Specification documentation	Poor, specification is still in development, however information regarding its development is available	Fair, specification documentation is not fully documented, however there is some amount of detail hosted online	Format lacks specification, but is well understood as a bit by bit copy of original material	Lacks standardization, describes a bit by bit copy of original disc, dependent on underlying file system
Backwards compatibility	Not backward compatible with previous AFF versions	Backward compatible, but deprecated	Yes, but dependent on underlying file systems to extract files	Yes, but dependent on underlying file systems to extract files
Descriptive metadata	Supports rdf-based assertions about the context of the stored disk image files and bytes.	Header contains a variety of information points regarding the host system that created the image	None	None
Cost to Support Imaging Environment	Low, python libraries are under development by the open source community	Moderate, original proprietor no longer distributes window based software, however open tools exist	Low, universal compatibility pending file systems	Low, universal compatibility pending file system

The above table explores some of the factors that might be considered when choosing to adopt a disk image format, as well as some examples of how various formats meet these criteria (This table is neither comprehensive or exhaustive. Many more factors and formats exist).

When selecting a disk image format for archival use, considerations include the openness of the format, backwards compatibility, sophistication of documentation, the ability for the format to capture descriptive metadata, the cost of supporting the imaging technology going forward, the adoption by consumer and professional markets, current archival adoption, etc.

The significance of each of these considerations will vary depending on the nature and scope of your need for a disk image. For long term archiving, backwards compatibility, cost, and open specifications reduce the risk of a format being rendered obsolete and unintelligible.

Disk Image format recommendations

Based on the above discussion of disk image formats, their uses, and the general archival goals and principles, we recommend the usage formats that have open specifications and good backwards compatibility. While the capacity of many proprietary formats to capture descriptive metadata might prove useful, they also provide a greater risk of obsolescence or abandonment, which is not desirable for a long term archival format.

Formats that meet our recommendation criteria include RAW (img, dd) disk images (for use when dealing with floppy disks or HDD/SSD storage) or .ISO when dealing with optical media. Both of these formats are bit level recreations of the cloned image that can be further processed or copied to derive access copies. They can be created with open source software (although we recommend also utilizing physical write blocking tools) and can be understood by any computer that recognizes the underlying file system of the disk image.

Further Reading

This section includes recommendations to readings that discuss topics in more detail or beyond the scope of this primer.

Disk Images:

https://en.wikipedia.org/wiki/Disk_image
A pair of blog posts discussing the process of utilizing the floppy disk imaging workflow at the University of Michigan's library:
https://www.lib.umich.edu/blogs/bits-and-pieces/disk-imaging-preservation-part-1
https://www.lib.umich.edu/blogs/bits-and-pieces/disk-imaging-preservation-part-2

Spreadsheet with advanced data on file format considerations (some data from this primer came from this spreadsheet):
https://docs.google.com/spreadsheets/d/18t-fU8ZO20Pgio6-QyPYHP3BR07lqL226gg4vRZPl84/edit#gid=1603236156

File Systems:

https://en.wikipedia.org/wiki/File_system

https://support.microsoft.com/en-us/help/100108/overview-of-fat-hpfs-and-ntfs-file-systems

Imaging Tools and Software:

Bitcurator project and Operating System:
https://bitcurator.net/bitcurator/

https://github.com/BitCurator/bitcurator-distro

Kryoflux:
https://github.com/archivistsguidetokryoflux/archivists-guide-to-kryoflux