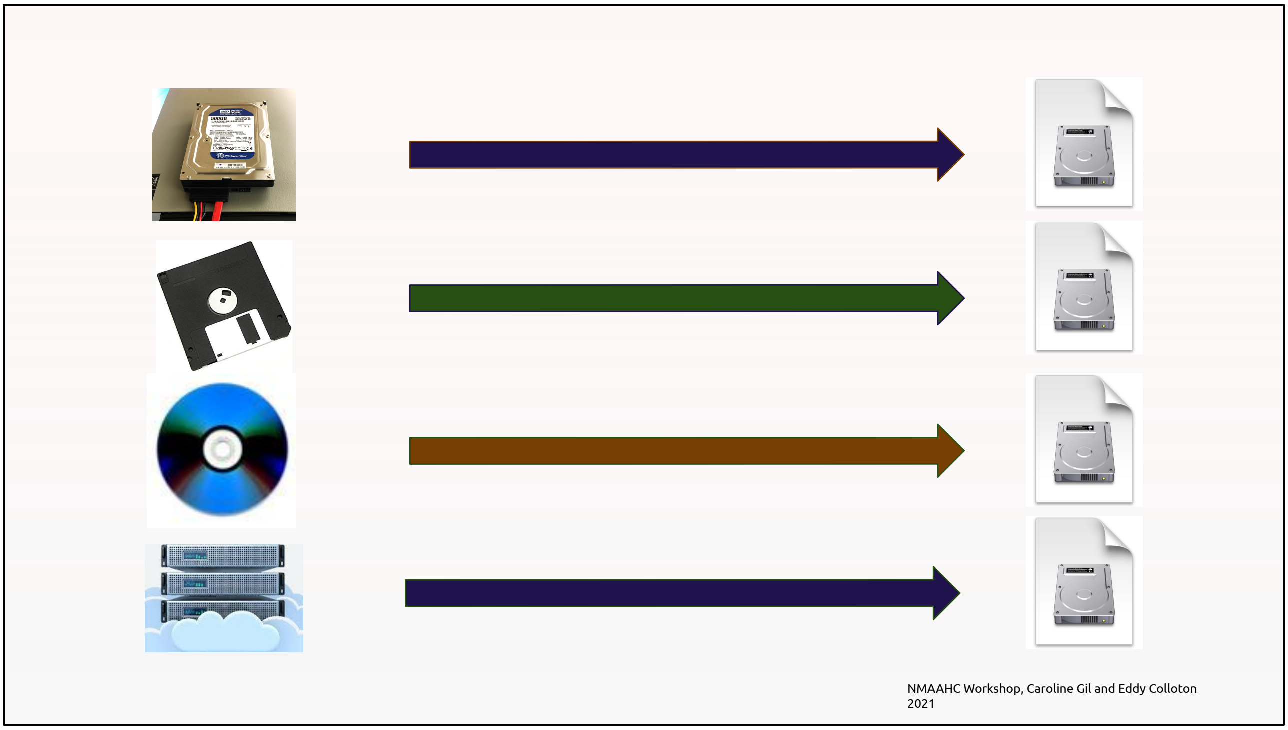
Disk Images

They’re just like us

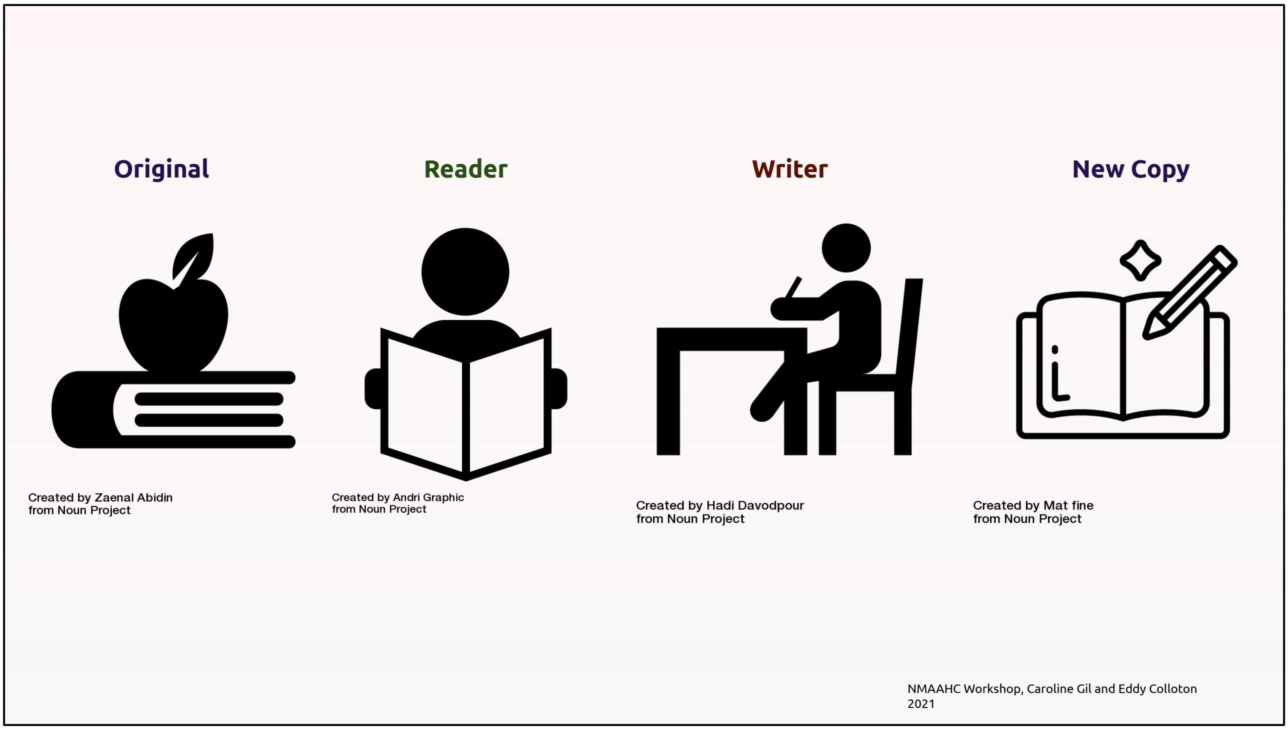
NMAAHC Workshop, Caroline Gil and Eddy Colloton
2021

We’re going to start at the beginning with some basic concepts about disk images and build up from there. As a part of this talk I’m going to introduce terminology and concepts that you may not be familiar with, feel free to ask questions in the chat or use the raise hand function, but also know that we’re going to unpack some vocabulary immediately following this.



NMAAHC Workshop, Caroline Gil and Eddy Colloton 2021

What is a disk image?
Wikipedia defines it as “a computer file containing the contents and structure of a disk volume or of an entire data storage device.” Which is actually a pretty good start. You can think of disk images as just a container. A file to store things in.
You can make disk images of lots of different things - a computer hard drive, a floppy disk, an optical disk, or just a partition on a server, any of these can then be wrapped up into a single file we call a disk image.



Original

Reader

Writer

New Copy

Created by Zaenal Abidin
from Noun Project

Created by Andri Graphic
from Noun Project

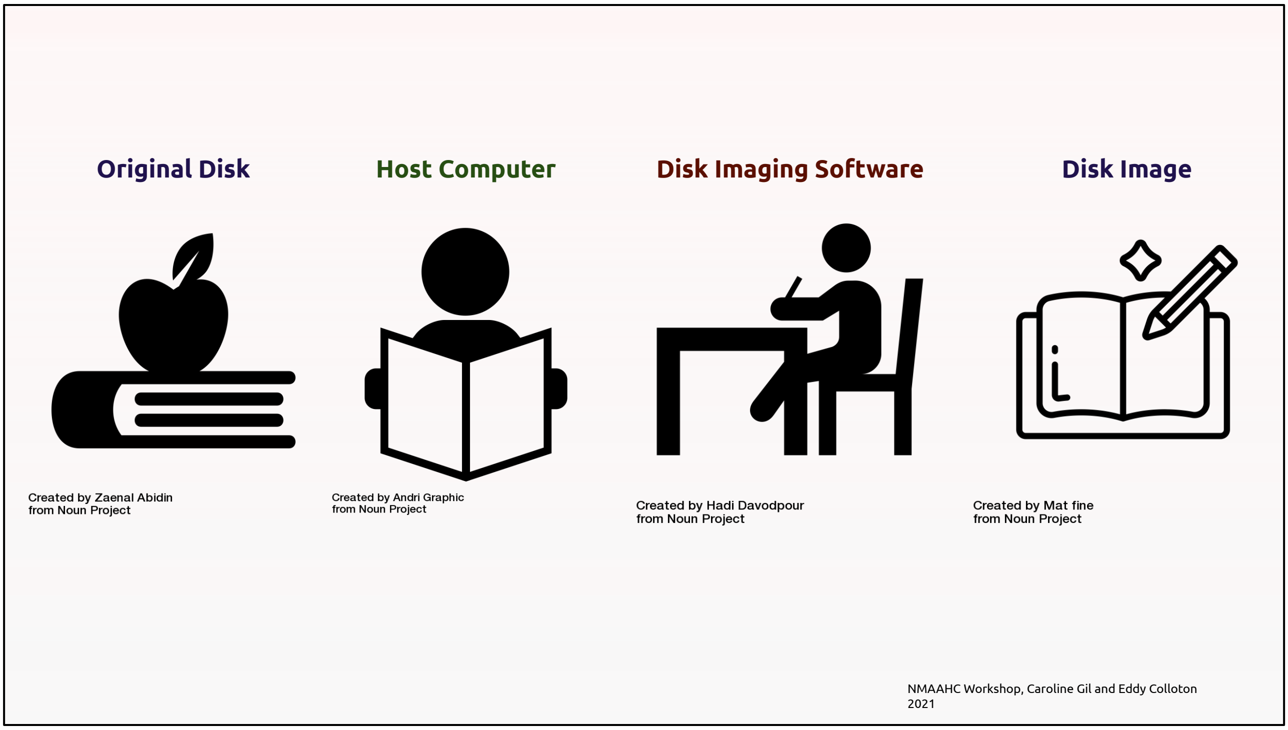
Created by Hadi Davodpour
from Noun Project

Created by Mat fine
from Noun Project

NMAAHC Workshop, Caroline Gil and Eddy Colloton
2021

I like to use an analog analogy when describing disk images. Comparing computer stuff to physical stuff always gets a little goofy, but hopefully this will be helpful. Think of a, admittedly very inefficient, methodology of copying a book by hand performed by two people. One person reads from the book, while the other person writes out everything the reader says into the book’s new copy. If there is a blank page in the book, the copier skips a page in the copy. If there is marginalia in the book, the copier mimics the marginalia perfectly. If someone has crossed out text in the book, the copy will contain the text, and include the fact that the text has been crossed out.

An important thing to remember is that if the reader omits a word, or accidentally skips a page, than the writer will still copy whatever the reader says, verbatim. Therefore the new copy will contain any errors that the reader makes. Similarly, if the original book is missing pages, or the text has become smudged, those errors will also be reproduced in the new copy.



Original Disk

Host Computer

Disk Imaging Software

Disk Image

Created by Zaenal Abidin
from Noun Project

Created by Andri Graphic
from Noun Project

Created by Hadi Davodpour
from Noun Project

Created by Mat fine
from Noun Project

NMAAHC Workshop, Caroline Gil and Eddy Colloton
2021

Some of you likely already see where I’m going with this. In my book analogy:

- The original book is the device or volume that you are imaging.
- The reader is your computer, which is reading the original volume. So if there are errors in reading the original volume, either due to damage to the original, or a bad read by the disk drive, then those errors are reproduced in the disk image.
- The Writer, the thing transferring data into the new copy, is the disk imaging software. There’s a variety of different disk imaging software, which we’ll discuss in detail
- The new copy is of course the disk image itself.

Types of Disk Images

- Physical Disk Image
- Logical Disk Image

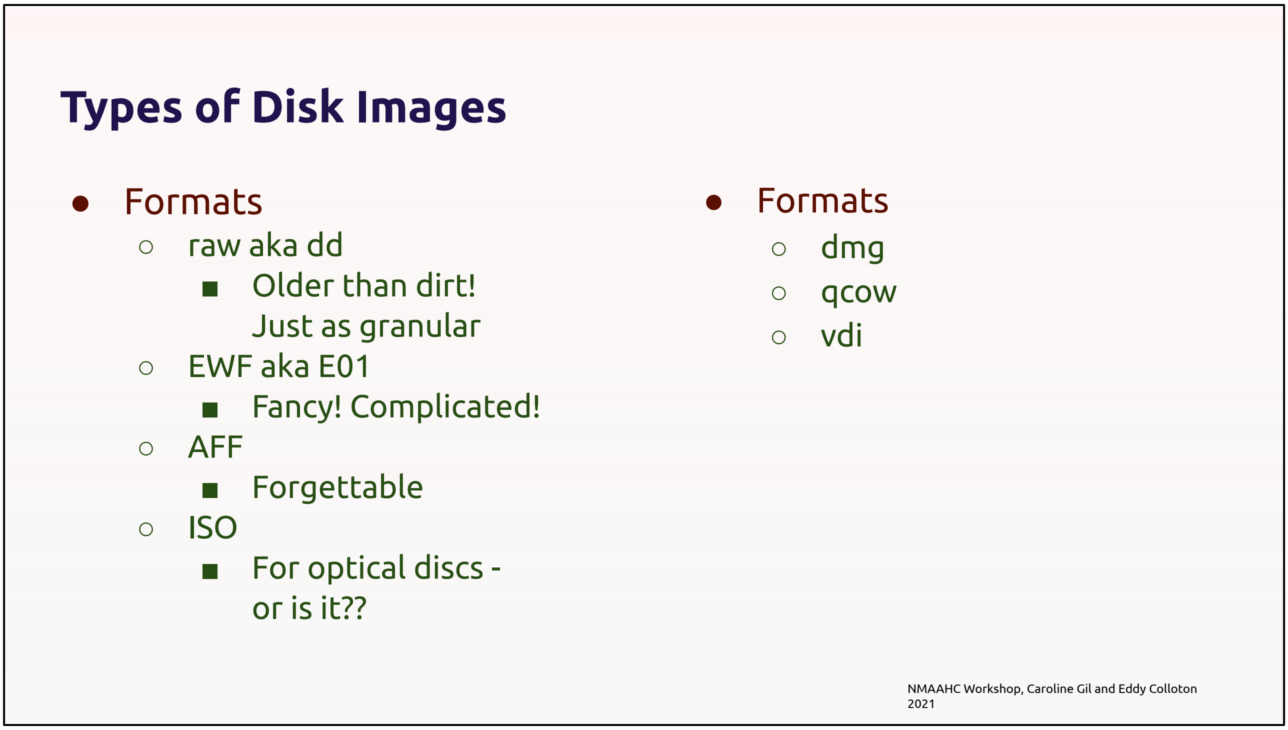
Created by Ben Davis
from Noun Project

Created by Srinivas Agra
from Noun Project

NMAAHC Workshop, Caroline Gil and Eddy Colloton
2021

There are different types of disk images. A physical disk image is what we create and encounter in cultural heritage most often. Throughout the workshop as we talk about disk images, this is the type we'll be discussing.
Simply put, a logical disk image only contains data a user would view if they were scrolling through a file explorer like "my computer" or "Finder." Whereas a physical disk image is a layer of abstraction above that. A physical disk image captures more data, including deleted files, and unpartitioned space. Telescopic difference between logical and physical disk images, physical disk images contain all the information within a logical disk image, and then some.

To be clear, these are not "formats" of disk images, they are "types." You're most likely to encounter this terminology - physical vs logical - when creating a disk image. There are a variety of different disk image formats out there, which I'm about to get into, but most formats can contain either logical or physical disk images.



Types of Disk Images

- Formats
  - raw aka dd
    - Older than dirt! Just as granular
  - EWF aka E01
    - Fancy! Complicated!
  - AFF
    - Forgettable
  - ISO
    - For optical discs - or is it??

- Formats
  - dmg
  - qcow
  - vdi

NMAAHC Workshop, Caroline Gil and Eddy Colloton 2021

The top two formats on the list on the left are the most common to cultural heritage, raw disk images and EWF disk images (sometimes called E01). For now I'll just say that raw disk images are completely uncompressed, they are an exact bit-for-bit copy of the target volume that was disk imaged, including empty space, and store no embedded metadata. The raw disk image format is basically as old as modern computers, created with the dd application in the 1970s. EWF disk images, sometimes called E01 are comparably new fangled, and are part of a broader category of "forensic disk images" they allow for compression, embedded metadata and encryption. We'll do a deep dive on the pros and cons of these two formats later on.

I also want to mention some other formats you might come across. AFF is a different kind of forensic image, which has lost favor to EWF. Although a new version AFFv4 may increase its popularity, only time will tell.

There is no comprehensive single specification for all of the variant formats called ISO image. They take their name from ISO 9660, a standardized optical disc file system, (more on file systems soon) but an iso disk image can contain any file system, and in fact often contain the other most common optical disk file system UDF. ISO is really just a file extension given to disk images. In fact, if you have a raw disk image with a file extension that your operating systems doesn't recognize, such as the ".001" or ".dd", simply changing the extension to ".iso" can allow your operating system to recognize and open the disk image. In this way, while associated with optical media, a disk image with an .iso file extension could be from any source.

The last three formats are all proprietary formats and are most commonly used as “virtual disk images” , intended to mount as attached file systems or serve as hard drives for virtual machines. The dmg, you likely recognize, as it is Apple’s proprietary format for delivering software. Qcow is the virtual disk image format for the emulation software qemu, and vdi is the virtual disk image for VirtualBox.

Compression!

Created by Eucalyp
from Noun Project

NMAAHC Workshop, Caroline Gil and Eddy Colloton
2021

I’ve mentioned compression so I want to touch on, in a very basic way, how that works. If we return the book analogy from earlier, you’ll remember that I mentioned that how all the little imperfections and inefficiencies of the original must be copied. A “blank page” of the book would result in a “blank page” in the copy. That is not the case with compressed disk images, like EWF disk images. In these disk images, instead of including a blank page, the copy might simply say “1 blank page.” In this way, compressed disk images contain the same amount of information without taking up as much room. They are losslessly compressed. When they encounter empty blocks on a target volume, they simply note their existence.

This is a concept we will return to but I want to make sure I have emphasized it - raw disk images will be the same size as the storage capacity of the volume that was disk imaged. A 500 GB hard drive with 2 GB of data on it will result in a raw disk image that is 500 GB. That’s 498 GB of empty space. A compressed disk image of the same hard drive will be much smaller.