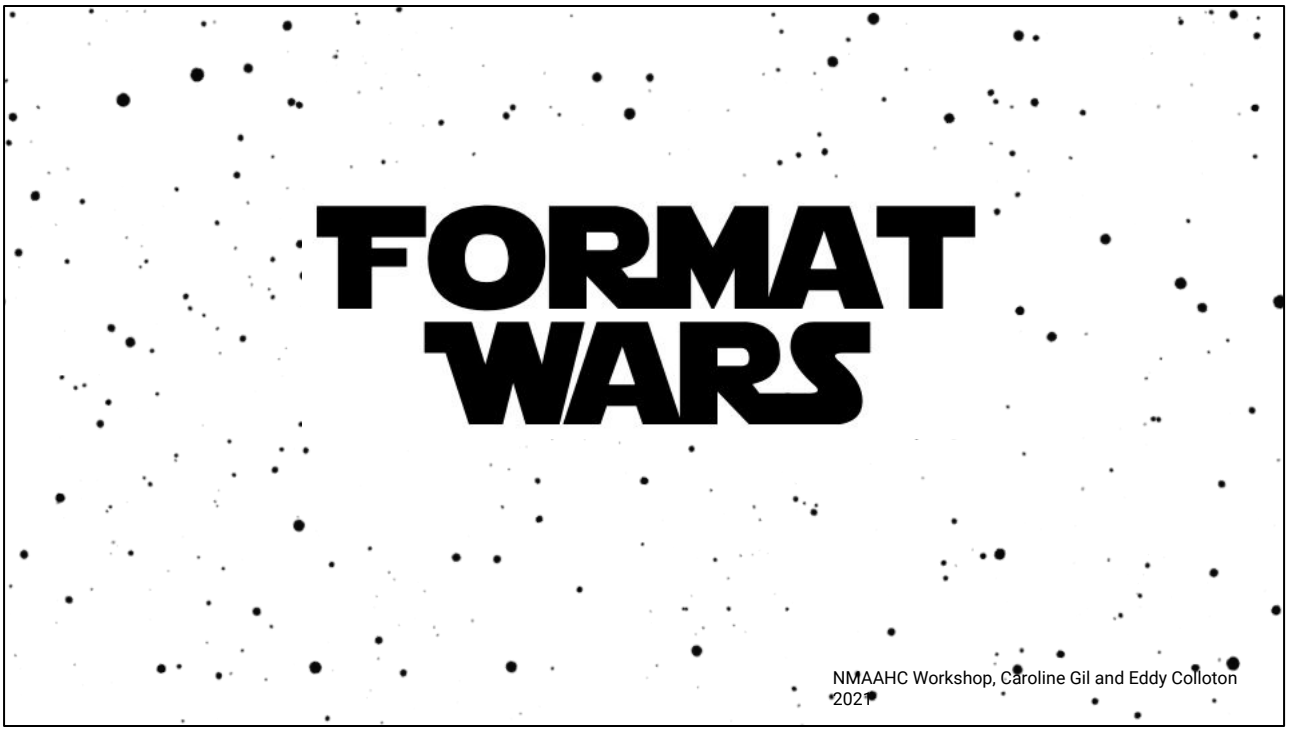
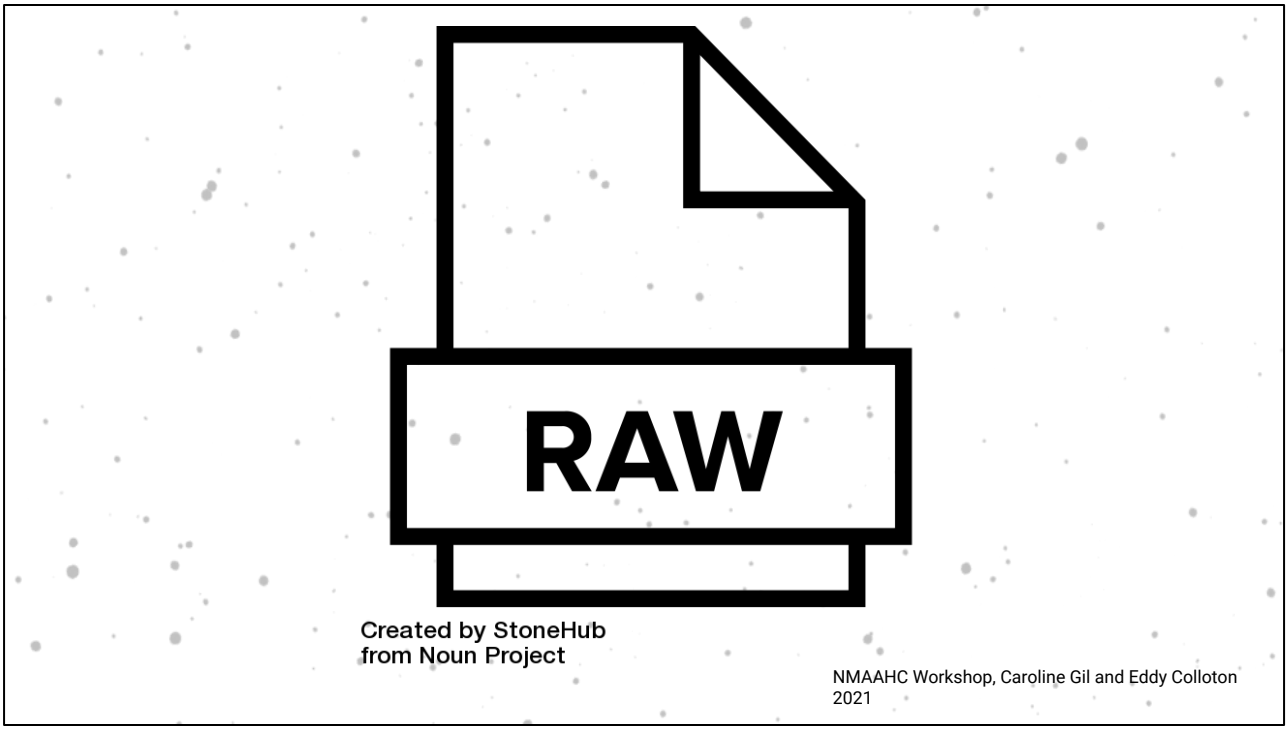


Like we've discussed, there are many different disk imaging formats. One format is not better than the other. The choice of which format to use should be context dependent, but the good news is that you have CHOICES. Your format choice will depend on a wide variety of factors that we'll discuss over the course of the workshop.



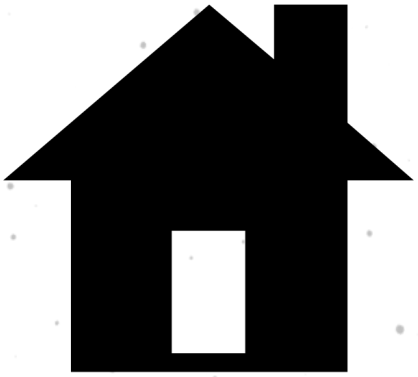
The two disk image format types that we're going to focus on are the two most common in cultural heritage.

A raw disk image is an uncompressed sector-by-sector sequence captured from a physical or logical volume. The raw image format is an open format, free of any license restrictions. Raw disk images can have any arbitrary file extension. Common extensions for raw disk images are .raw, .dd, or a numerical sequence such as .01. The "dd" file extension takes its name from a Unix command line application of the same name, used to copy and convert data (files or raw device contents) with a specified input and output

The Expert Witness Disk Image format, often referred to as E01, EWF, or the EnCase format, is a closed format defined by Guidance Software for use in their EnCase tool to store hard drive images and individual files. The EWF format is often called E01 because it allows for data to be stored across multiple segment files; these files are designated with sequentially numbered file extensions. For example; “.e01, .e02, e03” etc.

EWF Family

- EnCase 1-6
  - Subtle differences
  - disk imaging software does not distinguish between versions
  - Reverse engineered by Joachim Metz
- EnCase 7
  - aka EWF v 2
  - aka EX01
  - Not broadly adopted
    - FTK and Guymager do not make Disk images in this format
  - Also reverse engineered by Joachim Metz

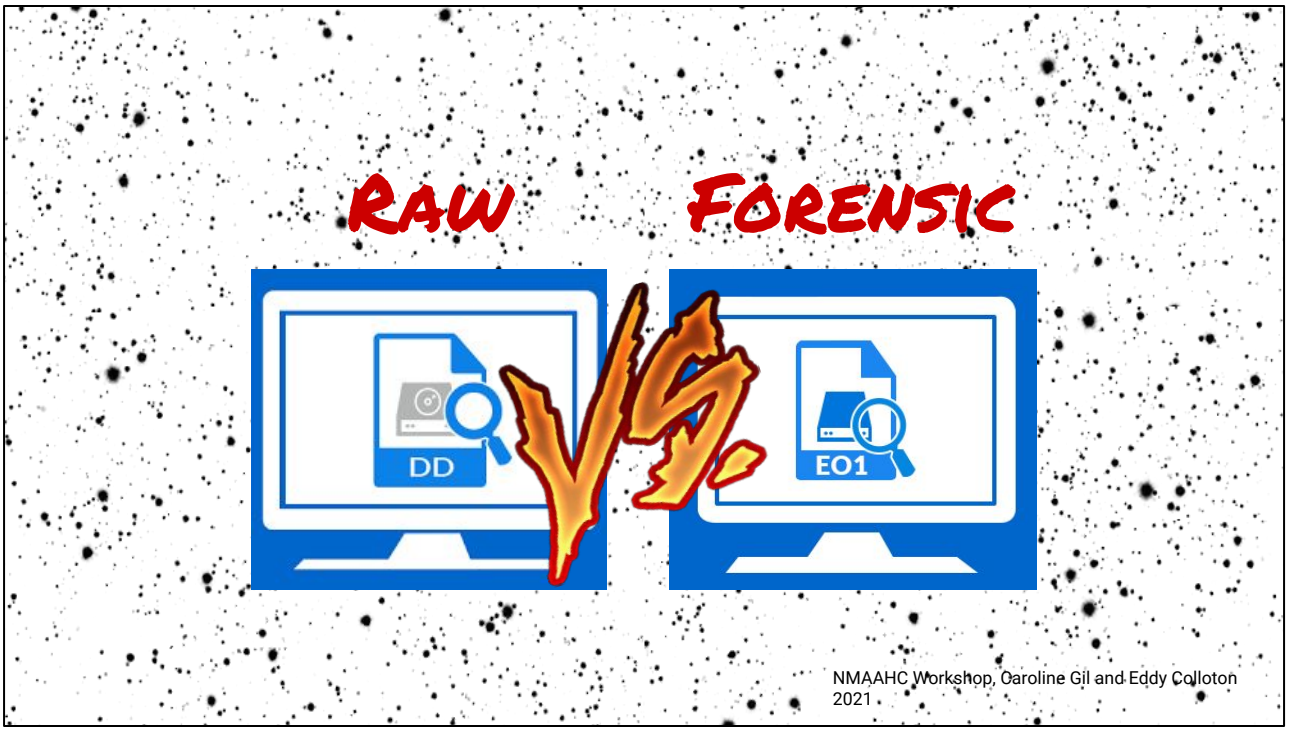


Created by Arief Sugiyanto from Noun Project

NMAAHC Workshop, Caroline Gil and Eddy Colloton 2021

EWF is technically a “family” of formats, there were several versions with very subtle differences, referred to as EnCase 1-6. The differences between these versions are relatively inconsequential. Tools for making disk images won’t even distinguish between them. This format was reverse engineered by Joachim Metz, and is published on github, so while it is proprietary it’s publicly documented.

This format was succeeded by a different format, most often referred to as EX01, or EWF version 2. However, this format is not widely adopted in archives, and the disk imaging software broadly adopted in our sector doesn’t even make this format. So for the purposes of our workshop, when we say EWF, we’re referring to the format that guymager, FTK, and libewf create, which have an E01 file extension. Metz has also reverse engineered this format, and published his findings on github.



Now that we've introduced the two formats lets compare the two

Considerations for a target file format

- **Compression**
  - Smaller File Size
  - Need to “decompress” with specific software
- **Metadata**
  - Beneficial for both description and for preservation
- **Sustainability**
  - Software dependency exposes the disk image to risk
  - Adoption
    - Adopted by whom? What are the community values?
  - Open source vs. Closed Source

NMAAHC Workshop, Caroline Gil and Eddy Colloton 2021

Criteria to help guide your decision will be:

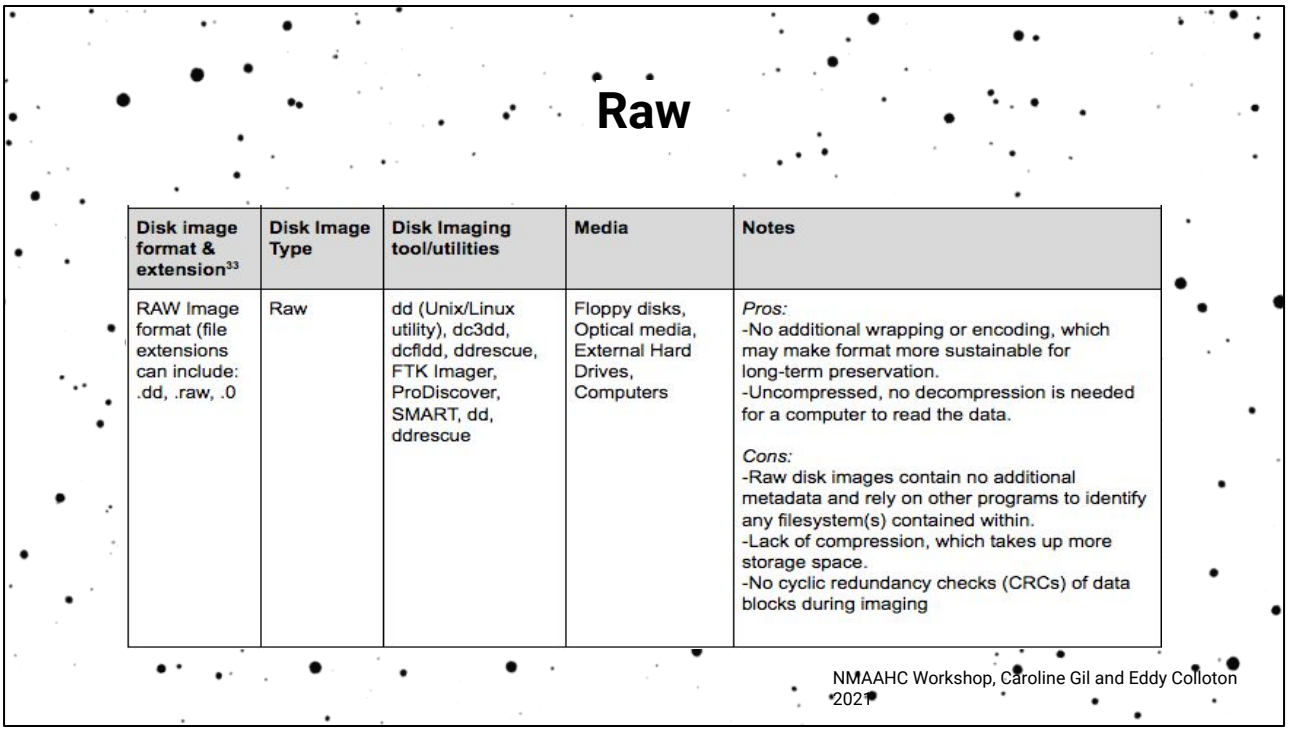
Compression

- Is the format compressed or uncompressed?
- VERY IMPORTANT TO NOTE that disk image formats only use lossless compression so there is NO issue of fidelity - no original data is lost when transferring between formats unlike with certain video compression formats
- Compressed formats take up significantly less space (perhaps up to 10 times less space depending on how much data is on the hard drive). In uncompressed formats, if you have a 10TB hard drive with only 1TB of data, you still have a 10TB disk image, a compressed format would produce a much smaller image.
- Compressed formats require software that can decompress the format
- Just like you need to have a codec installed in your video player to read video files encoded in a certain format, you need a the proper program capable of decompressing the data to read a compressed disk image

Metadata

- Does the format allow you to embed metadata in the disk image?
- Might want to embed descriptive information about what artwork or information about the computer that the disk image relates to, or information about who created it, when it was created, etc. etc.
- Some formats automatically embed preservation metadata like checksums and you can embed these within the file for certain formats

- the long term?
- How much do you want to risk on the software necessary for accessing the format surviving into the future?
- As we mentioned, we're repurposing formats developed by the digital forensics community, a community that is not thinking about long-term preservation and they may update or deprecate certain formats
- Is the format open source (format is essentially public domain), proprietary (the format is the intellectual property of one or several companies), or semi-open (format is technically proprietary but open-source tools have been developed for working with it)?
- How ubiquitous is the format? A more ubiquitous format is likely more sustainable than one that is not adopted widely



A Raw file is an sector-by-sector sequence captured from physical or logical volume, and contains the exact data as it existed in the source media without any additions or deletions. These file types require no additional wrapping or encoding

Raw disk images can have any arbitrary file extension. Common extensions for raw disk images are .raw, .img, dd, or a numerical sequence such as .01.

Disk Imaging tools: dd, ddresuce, FTK Imager, Guymaer, and many others

Commonly used with a variety of media including floppy disks, optical media, external hard drives, computer hard drives, etc.

Raw

PROS:

- No additional wrapping or encoding, which may make format more sustainable for long-term preservation.
- Uncompressed, no decompression is needed for a computer to read the data.

CONS:

- Lack of compression, which takes up more storage space.
- No cyclic redundancy checks (CRCs) of data blocks during imaging
- No metadata support

NMAAHC Workshop, Caroline Gil and Eddy Colloton 2021

The RAW Image format is an open format, free of any license restrictions. No additional wrapping or encoding is applied, which makes this format more sustainable for long-term preservation. But, because RAW files can't be compressed they tend to take up large amounts of storage space.

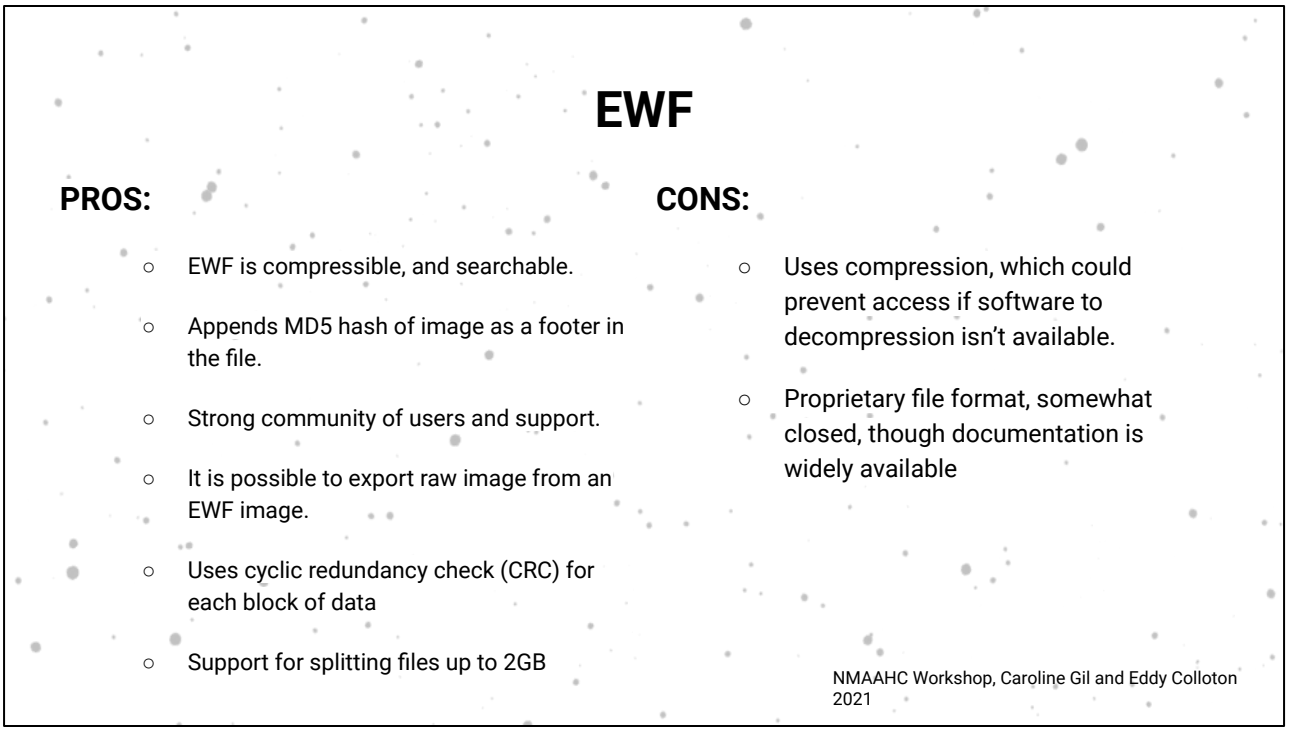
A disadvantage of this format is that it does not have the capability for embedded metadata and that it does not do any cyclic redundancy checks of data during the imaging process.

EWF

Disk image format & extension[32]	Disk Image Type	Disk Imaging tool/utilities	Media	Notes
EWF (.e01, or incrementing file extensions, i.e. .001, .002, .003 and so on)	Forensic, proprietary format.	Encase, FTK Imager (GUI and CLI), Guymager, libewf (ewfacquire). AFF, EnCase, FTK, SMART, Sleuth Kit, X-Ways can read this format.	Floppy disks, Optical Media, External Hard Drives, Computers	*Pros:* -EWF is compressible, and searchable. -Appends MD5 hash of image as a footer in the file. -Strong community of users and support. -It is possible to export raw image from an EWF image. -Uses cyclic redundancy check (CRC) for each block of data *Cons:* Uses compression, which could prevent access if software to decompression isn't available

As noted, file extensions is typically .e01, or for a sequence, .e01, e.02, .e03, etc.

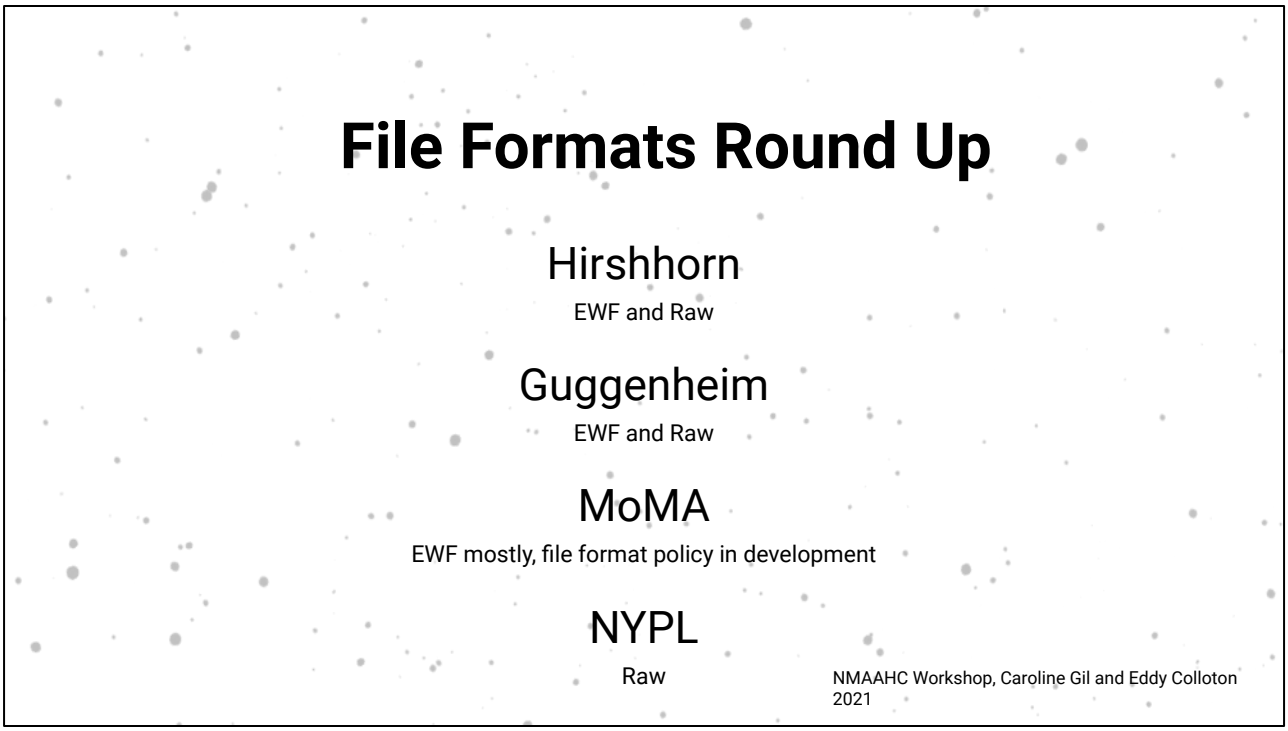
Software includes FTK, Guymager, and Libewf



One advantage of EWF, is that it applies Adler-32 checksums for every 64 block sectors, and MD5 checksums for the entire bitstream. Adler 32 is a faster version of CRC.

Allows for embedded metadata in the file.

The Expert Witness Disk Image format is a closed proprietary format. The future of the EWF format is unclear, (because it is a proprietary format made with the digital forensics community in mind. However, strong community of user support in cultural heritage as well, particularly in the archives and libraries fields, and the development of libewf by Joachim Metz, which resulted in an open-source tool to work with EWF files, are all good signs in terms of preservation.



As you may know, Caroline and I began our research on disk imaging as part of a collaboration between the Hirshhorn, the Guggenheim and the MoMA. These were the policies that we developed as a part of that project:

In the Guggenheim and Hirshhorn workflows, first we create EWF disk images, extract the raw disk images from the EWFs and store both. We've adopted this strategy because we like EWF for its built in file integrity checking, redundancy checks, every 64 blocks along with stored md5 or sha checksums for the entire bitstream. Plus there's the added benefit of being able to add embedded metadata which can be connected to entries in our collection management systems

However, since raw disk images contain no wrapping or encoding, they will likely be more sustainable. EWF is relatively new and has wide adoption, but among fairly niche communities. We don't know if software support will continue

By saving disk images in both formats in our repositories, hedging our bets. This is a pretty conservative strategy.. MoMA is still working towards establishing a disk image file format policy, though the media conservation department has already created EWF disk images, but also has some Raw disk images within their collection as well.

Caroline is of course now at NYPL, so I've added their disk image policy to the slide

as well, but obviously NYPL has a very different type of collection from modern art museums