# Disk Imaging Investigation

*Ashley Blewer on behalf of Artefactual Systems, prepared for the Denver Art Museum*

## Problem Overview

Denver Art Museum has a need to establish practices for evaluating extracted data from a disk image of a computer file system (as opposed to an optical disk or external storage file system) within the context of Archivematica.

Some outstanding questions at the start of this investigation:

* What problems are unique to a computer's disk image versus other forms of disk image, when the formats and processes are the same?

One issue with computer-based disk images are that the file structures of modern operating systems are notably more complex than external hard drives or optical disk images, resulting in hundreds of thousands of supplemental files that may or may not be relevant to the artwork seeking conservation treatment.

* How might the digital preservation community be able to benefit from using Archivematica to extract files from a disk image of a computer hard drive?

Anecdotally, institutions other than the Denver Art Museum do not extract disk images upon ingest into Archivematica. The Denver Art Museum chose to extract disk images when possible in order to create and store as much technical metadata as possible about the works for long-term preservation storage and to provide more granular data for future conservators to make more informed decisions when performing conservation treatment on the artworks, and better perform a more accurate representation of the original work when and if displayed again.

* What are the limits of Archivematica and other existing tools for sufficient analysis and extraction of complex disk images?

To summarize, Archivematica can have difficulty processing valid METS metadata for SIPs containing over 100,000 files, very large files can potentially cause problems, complex proprietary software can be difficult for metadata generation and file format analysis, and disk image formats may not be sufficiently self-describing for the tools within Archivematica to be able to diagnose and extract them after ingest. These points are expanded upon below, in the "known issues" section.

## Denver Art Museum Current Workflow

Documentation can be found here: <https://docs.google.com/document/d/1Z4LP8shWtdRYTkZrq1Q7nPomPWzJeoV369WC87kj0rM/edit> This documentation is still being drafted. The first few sections (which are more thoroughly drafted) focus primarily on the workflow designed and designated for optical disk images, not images consisting of external hard drives or operating systems. Documentation of the current workflow for computer hard drives is being written and added to the document now.

**[[ More info ? : Eddy Knows Best ]]**

**Tools**

- Guymager

- BitCurator

- Small environmental scan with FTK Toolkit (Level 5 Compression, chosen based on peer recommendation)

-SleuthKit

Tableau USB 3.0 Forensic Bridge

**Workflows**

*Optical disk images*: Create disk image using BitCurator. Ingest image into AM, have AM extract RAW image (other formats will not work)

*External drive disk images:* Same as above, depending on size. For larger disk images, the E01 disk image format was used, and all files were manually extracted pre-ingest using the tsk\_recover command with sector offset and block size determined from metadata acquired using SleuthKit.

*Computer disk images:* Under discussion below!

**[[ More info ? : Eddy Knows Best ]]**

## Known Issues

**Too many files**

Extracting a complex system results in hundreds of thousands of systems files that are being characterized and analyzed automatically by Archivematica's system, causing the process to move too slowly for production or halt completely.

**Overall file size is too big**

Importing can take a long time and possibly produce timeouts. Denver Art Museum's instance of Archivematica seems to have no trouble ingesting SIPs upwards of 200GB, so this is largely not a problem for disk images at this time. But slow uploading and ingesting time should be considered, especially in the case of selecting import methods in which significant data is included alongside the complete disk image in a SIP. Timeout limits may need to be adjusted to facilitate larger transfers.

**Disk image cannot be extracted after ingest**

Archivematica can only reliably handle RAW file formats (characterized typically by .dd or .img file extensions, but sometimes also the .iso format). The commands used by Archivematica do not have enough information to be able to sufficiently analyze and extract other file formats (.e01, for example) and therefore are not able to accurately extract files after ingest, as has been common with other disk images in Denver Art Museum's workflow.

**iPhoto Library**

Some imaged computers are able to be sufficiently ingested into Archivematica, but one problem manifested itself via the macOS's iPhoto Library, as this proprietary software produces an excessive amount of small files (upwards of 100,000) which can be too many for metadata to be sufficiently generated without the system timing out.

## Points of Failure

**METS generation failure**

The reason for this is that too much metadata is being generated and it causes the system to time out.

Failure mitigation:

**Disk image extraction within Archivematica**

Non-RAW disk image formats cannot extract ingested disk images due to lack of nuance in the SleuthKit command (/fpr/fpcommand/8f41dc6f-05eb-46d4-9b22-0a0d74673510/)

## Digital art assessment

When selecting significant files to ingest and generate metadata in Archivematica, it is important to develop a local methodology for digital content and assessment. This method must take into consideration the vast differences between digital artworks that have been acquired by the museum. It is important to note that this *must not* serve as a replacement to the ingest of a full disk image of the donated machine.

**Additional resources for policy development**

Archivematica Wiki on Disk Image Workflows (sponsored by UCLA Library and NYPL Special Collections): <https://wiki.archivematica.org/Improvements/Disk_Image_Preservation>

PERICLES ("The project aims to address the challenge of ensuring that digital content remains accessible in an environment that is subject to continual change"), <http://pericles-project.eu/>

Software Preservation Network: Collecting, Processing, and Providing Access to Software, <https://saaers.wordpress.com/2016/10/04/software-preservation-network-collecting-processing-and-providing-access-to-software/>

Tate: Authenticity, Change and Loss in the Conservation of Time-Based Media Installations, <https://www.tate.org.uk/research/publications/tate-papers/06/authenticity-change-and-loss-conservation-of-time-based-media-installations>

JISC project report.: The Significant Properties of Software: A Study, <http://purl.org/net/epubs/work/65878>

## Potential Workflows & Analysis

There are many options for the Denver Art Museum (and other institutions) for managing the preservation and long-term access of digital artworks, as preserved by full imaging of an artist-donated computer. Some viable options are listed below. It is not recommended that a conservator attempt to index and extract an entire harddrive, which would lead to problems indexing and metadata creation for too many files.

**Ingest disk image only**

*Ingest imaged disk without extraction.*

Pros: Image of computer remains preserved as imaged, ready for future conservation work to be applied. No decision-making, no extra arbitrary files.

Cons: No contextual information available to museum staff. No metadata generated for significant files. Entire disk image would have to be retrieved from storage and extracted manually. No context or information can be understood about the artwork without retrieving a copy from the archive and emulating the entire contents, which would also be unknown.

**Disk image and "Last modified date", significant date**

*Ingest a disk image and selected significant files from the disk image alongside. Gathering all files that seem to relate to an artwork based on a date chosen from knowledge gathered during the accessioning process.*

Pros: Easy access to significant files (videos, image) for museum staff of varying levels of technical expertise. Less files.

Cons: Duplicate materials, large file sizes. Inconsistent in selection, may not select a significant file.

**Disk image and "Last modified date", excluding system files**

*Ingest a disk image and selected significant files from the disk image alongside. Perform an investigation on the files found on the computer and collect everything.*

Pros: Easy access to significant files (videos, image) for museum staff of varying levels of technical expertise. Less files.

Cons: Duplicate materials, large file sizes. Inconsistent in selection, may not select a significant file. Hard to replicate across different types of artworks.

**Disk image and "user folder"**

*Collect all files in a user's "home directory."*

Pros:

Cons: Many files; more technical artists may be working outside of their user directory; still includes system files like default applications and unrelated applications

**Non-system portional ingest**

*Ingest only changed portions of a computer's file system representing the artwork alongside imaged disk.*

This option seems unlikely as a tangible solution to the current issue, even with allocated time and resources. Even if Archivematica or an independent system had the potential to contrast a file structure to a specific original operating system, the gains do not outweigh the risk that would come with lack of precision.

**Disk image with sidecar hierarchy/structure**

*This option is a proposal for, in addition to a folder containing a disk image and associated metadata files, the conservator to run a `tree` command to recursively list all files (More information: https://www.computerhope.com/unix/tree.htm) on a local, unpacked version of the disk image, and saving the directory structure of the entire image as a .txt file alongside the disk image and related metadata files. This is a slight upgrade from the suggestion of purely ingesting the disk image. Note that this command comes pre-installed on many Linux distributions but not natively available in macOS.*

Pros: Allows future non-technical experts to quickly see all files (with brief information and structure) that can be found in the image without having to extract the asset itself.

Cons: Running the command is a little bit of extra work for the conservator and can take some time (but not a significant amount of time). For example, running the command in the root directory of a modern and actively-in-use MacBook with a nearly-full 256GB hard\drive took 26 seconds.

## Other Institutions

Based on anecdotal, conversational inquiries into the workflows at other organizations, it seems that other institutions have not investigated this issue and, if ingesting disk images, do not attempt to extract them within Archivematica, and choose to submit them as they have been produced by their chosen disk imaging software (BitCurator, Guymager, FRED Forensic Workstations, etc.).

## Supplemental Information

It may be worth registering and storing supplemental metadata alongside ingested disk images. Harvard University's Disk Image Metadata Analysis Report (The report is available here: <https://wiki.harvard.edu/confluence/display/digitalpreservation/Disk+Image+Formats?preview=/204385887/212434673/02-01-DiskImageMetadataAnalysisReport.pdf>) recommends some other crucial metadata fields worth capturing when preserving disk images. This includes the format, version, size, date created, system format, bytes per sector, and sector count. This information can be gathered using The SleuthKit's `fiwalk` command. It's also important to contextualize the device on which the disk imaging process was performed. For the Denver Art Museum, this information may be stored in PREMIS metadata generated by BitCurator and could be packaged alongside the disk image when ingesting into Archivematica.