Digital Preservation Pipeline for Data Storage Media At The Cinémathèque Suisse

Imaging and extracting data and metadata from Special Collections media

|  |  |
| --- | --- |
| **Robin FRANÇOIS** | **Rebecca ROCHAT** |
| *Cinémathèque suisse**Switzerland**robin.francois@cinematheque.ch**0000-0001-5334-4438* | *Cinémathèque suisse**Switzerland**rebecca.rochat@cinematheque.ch**0000-0003-3908-3854* |
|  |  |

**Abstract – Bit rot and technical obsolescence are threatening the ability to read data storage media received by GLAM institutions. This paper presents the work in progress to build a pipeline of autonomous steps to correctly preserve information on data storage media. Inspired by video game and computer science preservation communities, our pipeline relies on promising open-source software, such as Aaru or HxCFloppyEmulator, and hardware such as Pauline. Challenges and limitations of our approach are discussed.**

**Keywords – data storage media, digital archival imaging, migration, obsolescence, media dump**

**Conference Topics – Community; Innovation**

# Introduction and rationale

The mission of the Cinémathèque suisse (Swiss National Film Archives) is to collect and preserve film archives, with an emphasis on Swiss cinematic heritage. Such collected archives are of a wide variety and include data storage media, which need to be processed and accessioned as any other acquisition. Archive collections have progressively collected these types of media, but cannot apply the same treatment as an analog collection. Alternative methods have to be explored.

Such data storage media, exclusively computer readable, cannot be handled by documentalists and archivists as straightforwardly as human-readable media. It requires specific equipment – which might not be easily acquired or used due to technological obsolescence – combined with technical skills that could pose a barrier for most staff members.

To streamline data storage media preservation, we have been building a pipeline. Analog media and digital audiovisual media (e.g., MiniDV, DVCAM, etc.) are excluded from this pipeline as they are handled by a separate team at the Cinémathèque suisse. This paper will focus solely on computer data storage media, such as floppy disks, optical disks or hard drives.

# Collection challenges

Data decay, also known as bit rot, can cause degradation to media that could complicate data reading [1]. Even if media can be correctly read, obsolete file systems and file formats can hinder any descriptive work. In addition, data storage media are data containers that can contain a very large amount of documents. The information stored on the ephemeral media should be copied and analyzed to prepare it both for accessioning and long-term archiving.

Data storage media are not human-readable and thus require a specific process for the necessary accessioning and descriptive work. Their longevity, even under good climatic conditions, has proven difficult to evaluate without extensive knowledge of the materials used [2] and the production process or the very precise examination of the media. This lack of detailed information has led most institutions to consider data storage media as fragile and in need of intervention before their end of life.

However, to unearth the information contained within, a reading device and specific software are required. Both elements are subject to obsolescence and access to the data is often prevented, not due to the conditions of the media but by technical and technological limitations.

To prevent a blind spot in the collections and a digital dark age, time is a key factor and imaging should be performed as soon as possible, before hardware and software become scarce and decay damages the media.

While a survey is still ongoing, more than 5,000 data storage media have already been identified at the Cinémathèque suisse, with mostly optical media (CD and DVD), 3.5” floppy disks and ZIP disks. It is expected that this figure will at least double once the survey is completed.

# Imaging Computer Data Storage Media

In order to properly preserve information on data storage media, the different data levels should be considered.

On digital media, the information is encoded as a sequence of binary values and stored using a physical property of the media (e.g., pits and lands for optical media). The devices required to read digital media will measure the variations in the physical property (the *signal*) and, through various operations performed by the device controller, generate a *bit stream*. The bit stream consists of *raw data* and control data (e.g., checksums, error corrections bytes, headers, etc.). Raw data can be interpreted as *user files and metadata* (e.g, modification date) through file systems and partitions [3].

As shown in Figure 1, the lower the level, the greater the amount of data, but the data is not always relevant. Accurately archiving media requires capturing all the necessary and meaningful data.



Figure 1 Levels of data

The level of necessary detail remains an open question, and it might be highly dependent on the institution and its missions. Initiatives such as DANNNG [4] are exploring and questioning the current situation.

Video game and computer science preservation communities have been approaching media imaging with great care and have developed methodologies and tools. These communities have been focusing on creating the most accurate images – sometimes to a very low level such as the signal level – and making the most of the processing in software *a posteriori*. This attention to detail has emerged due to two main factors: the importance of a near exact copy of the physical item and the copy protections that usually rely on peculiarities of the media.

# Choice of Hardware and Software Tools

As Aristotle reportedly wrote, “He who can do more, can do less”. Our pipeline was greatly inspired by the video game and computer science preservation communities, notably the Game Preservation Society [5] and the dumping.guide project [6].

As part of a GLAM institution, we also had to comply with our IT department to ensure that our software and hardware choices would integrate with a modern IT infrastructure. The priorities were acquiring brand-new equipment where possible and installing a supported operating system.

In order to fit some internal drives, 5.25" slots were needed. New tower workstations from major brands (e.g., HP, DELL, Lenovo, etc.) were evaluated. With limited choice, an HP Z2 G4 tower was selected as it covered our hardware requirements (connectors, storage, performances) and fitted with the IT department’s strategy.

As for the operating system of our imaging PC, our choice was a Linux distribution for its wide interoperability and large toolbox. As recommended by our IT department, we are using CentOS Linux.

For the drives, priority was given to new and internal drives with current connectivity (i.e. USB), rather than used drives and obsolete connectivity. The following list of drives covers most media found in the Cinémathèque suisse collections:

* Internal SATA ASUS BW-16D1HT optical drive, picked for its broad compatibility with optical media and the selected software suite;
* External USB Iomega ZIP 750 (used), capable of reading ZIP drives of all capacities;
* SONY MPF 920 3.5" floppy disk drive (new, manufactured in April 2009).

For controlling the 3.5” floppy disk drive, we have decided not to use KryoFlux [7] for various reasons: it is not a community-led, open-source project; the hardware relies on USB connectivity that is prone to errors; and the development of the project seems to have slowed in recent years. Instead, we have selected the open-source hardware and software project Pauline [8], developed by a consortium of non-profit organizations focusing on video game preservation (MO5.com, La Ludothèque française and the Game Preservation Society).

The Pauline daughter board, plugged into a Terasic DE10-Nano FPGA [9] running a Linux distribution, becomes a standalone hardware solution for reading the signal from floppy disk drives with Ethernet connectivity. The solution uses a small web server to give instructions and the resulting files can be retrieved via network share.

Signal files generated by Pauline are in the hxcstream format and can be opened with HxCFloppyEmulator [10]. Based on the recorded signal, the HxC software can reconstruct the bit stream, calculate checksums, identify reading errors or damaged sectors, and export the raw data as a raw image to be further analyzed with additional tools.



Figure 2 HxC floppy disk track analyzer. Red parts indicate sectors with errors.

For media other than floppy disks and to manipulate raw images, we rely on the open-source Aaru Data Preservation Suite [11]. Aaru’s philosophy is “to allow any user to create the best image (dump) that their hardware allows of the media they have” [12]. In addition, Aaru can be used to compare or convert existing images and list or extract files from an image with a broad range of supported image formats and file systems.

AaruFormat, Aaru’s own image format, allows lossless compression and deduplication but also contains image metadata, contents metadata and dump hardware information.

Since Aaru is central to our preservation pipeline, the Cinémathèque suisse has joined the Technical Committee of the Aaru Data Preservation Suite project to provide input and feedback on roadmap items and to ensure that the open-source project thrives.

# Preservation Pipeline

The main objectives of this pipeline are the following:

* Offload the archivists, documentalists, and the IT department of the digital archival imaging tasks
* Avoid the loss of fragile data on media due to physical and technological obsolescence
* Generate images of the media to allow the best possible software processing
* Provide the files and metadata to the archivists and documentalists to enable them to carry out their work
* Preserve and archive files in the best possible conditions to enrich digital collections and avoid the digital dark age



Figure 3 High-level view of the preservation pipeline for data storage media at the Cinémathèque suisse

Our preservation pipeline is summarized in Figure 3 in six steps.

The first step, “Find, inventory and catalog”, is very specific to our institution and its collections. Currently, data storage media have not always been identified and described, and are not often separated from the rest of the “analog” collections. Ongoing work consists of improving the handling of new acquisitions and applying the new procedures retroactively to the entirety of the Special Collections. Regrouping data storage media would allow batch processing, as well as improved conservation measures and packing. Attributing a unique identifier to each media is essential for further processing and associating the data extracted from the media to the right context.

The second step, “Identify the type of media and prepare for reading”, is necessary to address two issues. Media imaging requests might come from other teams that do not have full knowledge of our imaging capabilities, and some requests might require drives or tools that we do not have. In this case, an assessment will be made as to whether acquiring the necessary equipment and testing the methodology is relevant or if the imaging process is to be outsourced. The second issue is that precise media identification could need expertise, tools, and time that archivists and documentalists might not have. Cleaning and taking a picture of the media is also part of this step.

The third step, “Generate an image of the media”, consists of imaging the media at the lowest level necessary for adequate preservation, but also at the lowest level possible with the available drives and tools. In the current state of media drives, signal level imaging is only attainable for floppy disks. For other data storage media, bit-stream level imaging is the usual limit of the drives, and raw level imaging is often satisfactory for most situations. Capturing metadata about the media and tools (paradata) is also performed at this step.

The philosophy of this step is very similar to digitization: create a digital surrogate of the media with the best accuracy and quality. Any intellectual processing can be performed in separate steps.

Once the best image possible has been made from the media, the fourth step is “Extract files and metadata from the image”. Images are rarely directly useful for archivists and documentalists to perform their work. Files and folders need to be extracted from the partitions and file systems stored on the image.

Interestingly, this step can be performed in a decoupled way from the imaging step and can be automated or performed by a separate staff member. We are using Aaru for this step, except for file systems not yet supported (see Chap. VI.).

Once files, folders, and metadata have been extracted from the image, the process of accessioning and digital archiving (following the OAIS model) begins: files need to be converted (normalized) to allow archivists and documentalists to sort, accession, and describe the documents that will be added to the digital collections and the OAIS archive.

The last step of the pipeline is the archiving of the selected documents according to the OAIS model. The description of this step is beyond the scope of this article.

# Challenges and Limitations

The Cinémathèque suisse is still investigating the best approach for appraising “digital bulk” extracted from some media, such as the personal 2TB hard drive belonging to a film director. Generally, extracting files from media can be a technical challenge, but adapting existing appraisal processes to digital assets is always a challenge.

Mass digitization is still being evaluated, notably due to the large amount of optical media in our Special Collections. The Acronova Nimbie USB Plus automated loading system [13] is being considered, but compatibility with our current tools and pipeline needs to be tested.

Furthermore, the Aaru software has several limitations:

* Aaru is mainly a command line tool, which limits its use by less technical users.
* Imaging of large disks in AaruFormat is suboptimal due to bugs or missing functionalities. Imaging in QCOW2 is a good workaround while AaruFormatv2 is developed.
* The Cinémathèque suisse receives numerous media created in an Apple environment, with specific file systems. Aaru is currently not able to extract files from HFS and HFS+ file systems, which forces us to turn to a more manual approach using standard Linux tools.
* Metadata currently generated by Aaru needs to be transformed to fit into METS and PREMIS metadata used for digital archiving.

These limitations have been communicated to the Aaru project and should be addressed in future releases.

#### ACKNOWLEDGMENT

We thank our colleagues from the Cinémathèque suisse and the dedicated communities and institutions developing and supporting the Aaru, HxC and Pauline open-source projects.

# REFERENCES

1. M. Baker et al., "A fresh look at the reliability of long-term digital storage", *SIGOPS Oper. Syst.* Rev. 40, 2006, pp. 221–234.
2. J. Iraci, “Longevity of Recordable CDs, DVDs and Blu-rays”, Canadian Conservation Institute (CCI) Notes 19/1, 2019
3. B. Hayes, "Bit rot", *American Scientist*, 1998, vol. 86, no 5, pp. 410-415.
4. Digital Archival traNsfer, iNgest, and packagiNg Group, https://dannng.github.io
5. Game Preservation Society - https://www.gamepres.org
6. Dumping guide project by Hit-Save! - https://dumping.guide
7. Kryoflux - https://www.kryoflux.com
8. Pauline project webpage
https://www.laludotheque.fr/projets-en-cours/preservation-des-disquettes-pauline
9. Terasic DE10-Nano FPGA product webpage
https://www.terasic.com.tw/cgi-bin/page/archive.pl?No=1046
10. HxCFloppyEmulator, https://sourceforge.net/p/hxcfloppyemu/
11. Aaru Data Preservation Suite, https://www.aaru.app
12. N. Portillo, “Why is Aaru different?”, Blog post, 24 July 2020
https://blog.claunia.com/why-is-aaru-different/
13. Acronova Nimbie USB Plus product webpage,
https://disc.acronova.com/product/auto-blu-ray-duplicator-publisher-ripper-nimbie-usb-nb21/9/review.html