Digital Preservation Guide: 3.5-Inch Floppy Disks

Caralie Heinrichs and Emilie Vandal

ISI 6354

University of Ottawa

Jada Watson

Friday, December 13, 2019

Table of Contents

Introduction

The floppy disk was once ubiquitous. More than five billion were sold per year worldwide at its peak in the mid-nineties. Floppy disks helped enable the "personal computer" revolution and the creation of an independent software industry that now includes more than 10,000 companies. With the last company ceasing production of floppy disks in 2011, the diskettes have quickly become obsolete (International Business Machines Corporation, n.d.). It is becoming increasingly difficult to retrieve disk contents as the necessary software and hardware disappear, leading the archives community to borrow techniques and tools from the field of digital forensics, notably the practice of disk imaging. A disk image is an exact copy of the original disk, permitting archivists to safely appraise and manipulate files without compromising the original, while providing a digital copy for preservation.

However, the process of digitizing and preserving floppy disks is not without its challenges. After presenting a history of the floppy disk, the challenges related to digitization and preservation is explored. The following section presents the best practices that emerge from the literature related to digital archives, including the tools most commonly used. The final section proposes a workflow to digitize and preserve 3.5-inch floppy disks at the University of Ottawa Archives and Special Collections (ARCS).

History of the Floppy Disk

Where, when, and by whom was it developed?

According to Oxford University, a floppy disk, "also called a floppy, diskette, or just disk, is a type of disk storage composed of a disk of thin and flexible magnetic storage medium, sealed in a rectangular plastic enclosure lined with fabric that removes dust particles. Floppy disks are read and written by a floppy disk drive (FDD)" (Disha Experts, 2019). Its name comes from the 5.25-inch floppy disk, which was dubbed "floppy" because of its disk packaging of a very flexible envelope. Even though the 3.5-inch diskette was held by a rigid case, the name continued to prevail (Brown, 2001).

The floppy disk project was first started at International Business Machines Corporation (IBM) data storage location in San Jose, California. In 1967, a small team of engineers under the leadership of David L. Noble, including Alan F. Shugart, Herbert Thompson, Ralph Flores and Warren L. Dalziel, was tasked with developing a reliable and inexpensive system for installing software on computers. There was also a need to find an easy, convenient and less expensive way to send software update to their consumers. IBM started to sell floppy disks in 1971 and, in 1972, they received U.S. patents for the floppy disk drive and the diskette (International Business Machines Corporation, n.d.; Vaughan-Nichols, 2017).

The big storage breakthrough came in 1977 when Apple first introduced the Apple II. This first mass-produced computer came with two 5-1/4-inch floppy disk drives. After this, it became standard to have two drives in personal computers. IBM made floppy disks for many years and it continued to innovate up to the 3.5-inch floppy disk that became standard in the computing world. For example, the first 8-inch diskette had a storage space of 80 kilobytes and by 1986, the 3.5-inch disk had 1.44 megabytes storage capacity. The newest floppy disk was also more convenient because of its size and had a sturdier construction compared to previous generations (International Business Machines Corporation, n.d.; Vaughan-Nichols, 2017).

In the 1990s, 3.5-inch floppy disks were widely popular, but with the arrival of read-writable CDs, DVDs, and the USB flash drive, diskettes quickly became obsolete. It took a decade before they disappeared, and in 2011, Sony, the last company manufacturing them, announced their discontinuation. In response to this news, BBC News Magazine asked how their readers used floppy disk today. There were still many that used diskettes to update or back up old systems, but others found different uses: some used them as drink coasters, spatulas and clothing accessories (International Business Machines Corporation, n.d.; Vaughan-Nichols, 2017).

Why was it developed?

The team led by Noble was asked to develop a "reliable and inexpensive system for loading microcode into the IBM System/370 mainframes using a process called Initial Control Program Load (ICPL)" (International Business Machines Corporation, n.d.). This project was created to replace the commonly used tape drive and the result was the creation of the 8-inch floppy disk. By developing the diskette, IBM put aside the punch cards and showed their willingness to adapt to the evolving technology.

The floppy disk was a major step forward in user-friendliness for individuals. Instead of being tied down to one computer, a student could work at the library, save their work on a diskette, and continue working later at home on the same project. It was easy to transfer data from one computer to the other, but perhaps the greatest impact of the diskette was in the IT industry (International Business Machines Corporation, n.d.). Thanks to the floppy disk, instead of writing programs on individual computers, companies could now write programs, put them on disks and sell them through mail or in stores: "It made it possible to have a software industry" (International Business Machines Corporation, n.d.).

How Does a 3.5-inch Floppy Disk Work?

The 3.5-inch floppy disk can easily be compared to a cassette tape. The floppy disk and cassette tape use a thin plastic base material coated with magnetic iron oxide on both sides. Both can record information immediately, erase and write data, and can be very inexpensive and easy to use. One of the major differences from a cassette tape is its shape. While the cassette tape's magnetic tape is a long thin ribbon, the floppy is arranged like a disk. In this way, the floppy disk is similar to a vinyl record: the tracks are arranged in concentric rings, also called sectors, which permit the software to jump from file to file without needing to go through the files in between. The read/write heads move to the correct track, providing direct access storage (Brown, 2001).

Major parts of a floppy disk

When taking apart a 3.5-inch floppy disk, you will find two coloured plastic squares: the housing that holds the other smaller parts of the diskette, also known as the disk jacket. Below is a brief explanation of the major parts of a 3.5-inch floppy disk, accompanied by Figure 1 and Figure 2:

1. **Paper ring:** The magnetic disk is between the two paper rings. They are glued to the plastic housing and they do not move while the floppy disk spins. They also clean the disk by removing microscopic bits of dust (The Science Explorer, n.d.).

2. **Hub:** The hub is the metal centre of the magnetic disk that holds the disk together while it spins. (The Science Explorer, n.d.)

3. **Magnetic disk:** The disk is covered with iron oxide that can be magnetized. When data is saved on the floppy disk, a recording head creates a magnetic pattern on the iron oxide in a form that the computer can read and access the data anytime (The Science Explorer, n.d.).

4. **Shutter:** The shutter is folded over one edge of the disk which goes inside the computer first. Inside the computer, the shutter slides over and the data can be read through the rectangular slot (The Science Explorer, n.d.).

5. **Spring:** When the floppy disk comes out of the computer, the spring shuts the shutter and prevents any dust or fingerprints to get on the magnetic disk (The Science Explorer, n.d.).

6. **Write protect tab:** It is in the upper right corner of the diskette. When the hole is slid open, the floppy disk is locked and the computer will not allow you to add or erase

anything (The Science Explorer, n.d.).

7. **Plastic flap:** It is located under one of the paper rings and acts as a spring to push down the paper ring as closely as possible to the magnetic disk. (The Science Explorer, n.d.)

8. **Read/Write Head:** It is located on both sides of the floppy disk. The heads are not directly opposite of each other in an effort to avoid interaction between write operations on each surface. One head is used for reading and writing while the wider head is used to erase a track prior to begin writing (Brown, 2001).

9. **Stepper Motor:** This motor makes a number step to move the read/write heads to the proper track position (Brown, 2001).

10. **Floppy disk drive:** A small spindle motor engages the hub, spinning it at 300 or 360 rotations per minute (RPM) (Brown, 2001).

11. **Boot Sector:** The boot sector is the area where the Master Boot Record (MBR) is located; it is the first sector of the disk. The MBR is the program that runs when the computer starts (Indiana University, 2018).

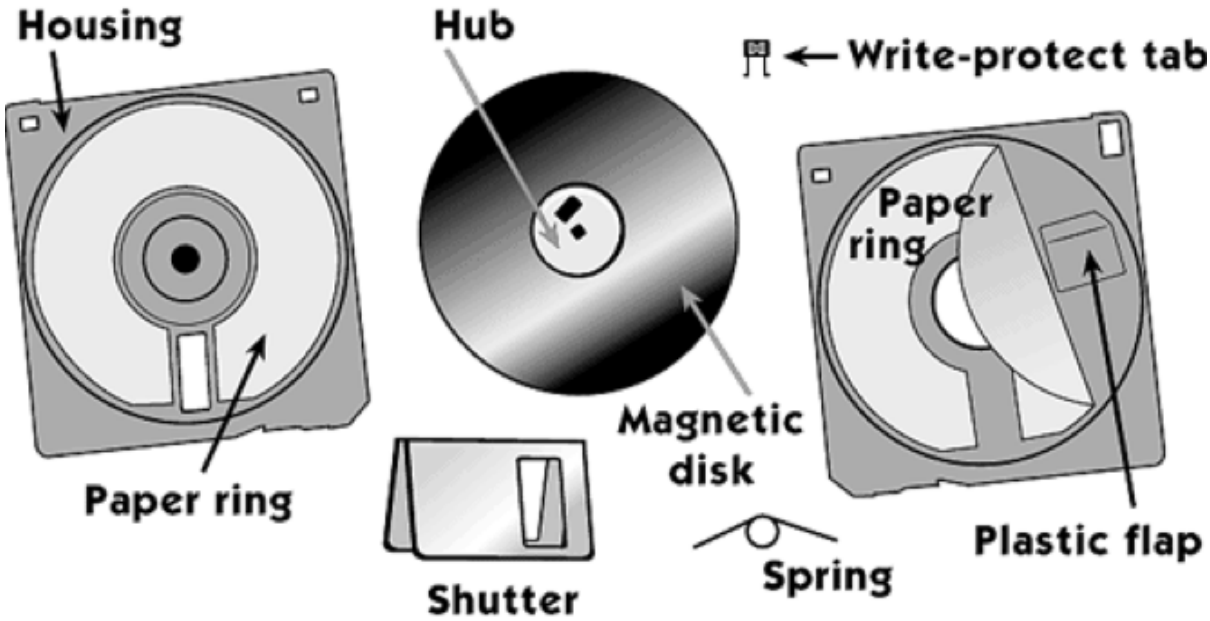

Figure 1. What's inside a floppy disk. From the Science Explorer. (n.d.). Dissect a Disk – Find out what's inside a floppy disk [digital image]. Retrieved from https://www.exploratorium.edu/science_explorer/dissect_disk.html

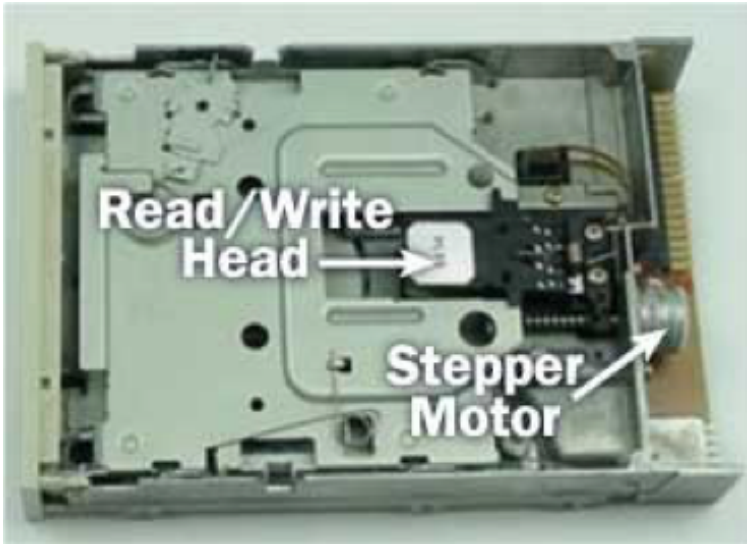

Figure 2. Floppy Disk Drive Terminology. From Brown, G. (2001). How Floppy Drives Work. Retrieved from https://computer.howstuffworks.com/floppy-disk-drive1.htm

Writing data on a floppy disk

There are many steps involved in writing data on a floppy disk, with similar steps involved when reading data on a diskette (Brown, 2001):

1. First, the computer program instructs the computer hardware to write data on a floppy disk.
2. Second, the computer hardware and the floppy disk drive start the motor in the disk drive to spin the disk.
3. Third, the stepper motor starts working to access the correct track. The floppy disk drive electronics is formatted to determine how many steps the stepper motor needs to turn to move the read/write heads to the correct track.
4. Fourth, the read/write heads stop at the correct track and the read head checks the pre-written address on the floppy disk.
5. Fifth, before the data is written to the diskette, an erase coil from the read/write head is activated to clear a clean slate before the write head is enabled. The erased sector is created in order to have no interference from sectors in adjacent tracks.
6. Sixth, the write head starts working and puts new data on the floppy disk "by magnetizing minute, iron, bar-magnet particles embedded in the diskette surface" (Brown, 2001).
7. Finally, the floppy disk stops spinning and waits for a new command.

Preservation and Digitization Challenges

Physical damage and degradation

Digital archives literature and digital forensics literature focuses on disk imaging (creating a sector by sector copy of the entire disk) as a solution to transfer information from obsolete floppy disks to contemporary formats for storage and access (Durno, 2016; Lee, Woods, Kirschenbaum, & Chassanoff, 2013; Waugh, 2014). In order to create these disk images, floppy disks need to be kept in good physical condition to be read. As with any physical media, the way that floppy disks are stored and handled can have a significant impact on their condition. Physical damage caused by improper care as well as natural degradation over time pose a threat to the archivist's ability to preserve and digitize diskettes.

According to the Canadian Conservation Institute, floppy disks should be stored in a cool, dry place below 23°C and below 50% relative humidity, as hotter and more humid conditions reduce the lifespan of the disks (Dicks & Iraci, 2017). High levels of heat or humidity can cause the binder glue that holds the magnetic particles on a disk to either become soft and sticky, or brittle, causing the magnetic particles containing recorded information to flake off. If left in direct sun, high heat may physically warp or shrink the media making it impossible to read (Andolsen, 2006; Gough, 2013; LeFurgy, 2012; National Archives of Australia (NAA), n.d.).

Clean storage cases should be used to keep dust and debris away from the disks (Dicks & Iraci, 2017). Particulate matter can scratch the magnetic media, resulting in permanent loss of disk contents. Over time, dust or debris may adhere to the surface of the disk causing read errors, such as causing the head to skip along the surface of the disk (Gough, 2013). Likewise, touching the surface of the disk can leave damaging fingerprints (Ahl, 1983; Dicks & Iraci 2017; NAA, n.d.). In these cases, the disk itself may be too damaged to attempt the imaging required for digitization (Durno, 2016).

Though the plastic disk jacket of 3.5-inch floppies affords some protection, care should be taken to not bend or twist disks (Ahl, 1983; Dicks & Iraci, 2017; NAA, n.d.). Storing disks on top of each other, jamming into boxes, or placing heavy objects on top of them can cause warping in the jackets making the disk spin unevenly, resulting in erratic reading and writing (Ahl, 1983; NAA, n.d.). If only the jacket is damaged, cutting it open and transferring the disk to a new undamaged case may be a viable solution (Durno, 2016).

Even the most well stored disks are susceptible to bit rot (Gough, 2013; McKinley, 2015, Waugh, 2014). Over time, "the magnetic media slowly demagnetizes itself due to stresses like thermal cycling and alternating magnetic fields nearby or even adjacent disks and movement of those disks" (Gough, 2013). As bits lose their magnetic orientation, data on the disk becomes corrupt (Salter, 2014; Wikipedia Contributors, 2019). Demagnetization can also occur

instantaneously as a result of exposure to a magnetic field (Andolsen, 2006; Ahl, 1983; NAA, n.d.).

Given the vulnerability of floppy disks to physical damage and degradation, transferring data off floppy disks is key for its preservation (Waugh, 2014). It is important to recognize that time is a significant factor affecting the ability to access content stored on floppy disks, as they have a relatively short viable lifespan (Andolsen, 2006; Gough, 2013; McKinley, 2015). Estimates of their longevity vary depending on the source; Dicks & Iraci (2017) of the Canadian Conservation Institute cite five to fifteen years, while other companies like Verbatim and Maxwell suggest thirty years or more (Ahl, 1983). However, the lifetime of the media as it depends on physical properties and conditions is only one aspect of the longevity of floppy disks.

Hardware and software obsolescence

The obsolescence of hardware and software required to access floppy disks is a serious challenge for their preservation and digitization (Dicks & Iraci; McKinley, 2015). While using legacy hardware (such as disk drives and computers) and operating systems is an option for imaging floppy disks, the majority of literature surveyed used contemporary software and operating systems, sometimes in combination with older disk drives. Legacy technology is often difficult to acquire, time consuming and complicated to set up, and difficult to use (Waugh, 2014). While a legacy machine with an internal disk drive may read certain disk formats very well, it lacks the flexibility of modern tools designed for digital archivists that can analyze the disk format and recover the data, as well as presenting additional challenges such as the lack of modern ports to export data (Gough, 2013; Waugh, 2014).

Identifying the formats of floppy disks that need to be imaged is a challenge that can prevent the disk from being read and imaged. In Emory University's experiments with seven different approaches to imaging floppy disks, Waugh (2014) notes that "the formatting of the disk appeared ultimately to have the greatest significance in determining whether or not an image could be captured." Durno (2016) explains how competing computer system developers from the 1980s implemented many proprietary technologies such as disk encodings, file systems, and file formats, making digital information far more enmeshed in the particular technologies that created them than is the case today.

Both Durno (2016) and Waugh (2014) acknowledge that there is a certain amount of trial and error required when attempting to read floppy disks due to the wide range of formats. The hardware and software needed to read and image the disks depends on the format of the media (McKinley, 2016). While imaging software, such as that of BitCurator and KryoFlux, comes with built-in formatting settings for known disk types, certain formats are not easily identified.

Though in some cases relevant information is found on the disk label, determining what combination of hardware and operating system was used to write the disk is not always obvious and requires research and some trial and error.

Certain disk formats are not readable by standard USB disk drives, such as Mac-formatted disks, Microsoft's DMF format with varying sector lengths, and copy protected disks (Durno, 2016; Gough, 2013; Waugh, 2014). To prevent unwanted copies of disks, software companies implemented copy protection measures which can still be in effect years later as archivists attempt to copy the disk for preservation purposes. As such, a disk with many imaging errors may not be a damaged floppy disk, but one that is copy protected (Archive Team, 2018). Using the KryoFlux has been a successful work around for such disks, as well as for reading "flippy disks" (single-sided floppies with data written on both sides) in combination with a modified drive (Durno, 2016; Gough, 2013; Waugh, 2014; Archive Team, 2018).

The physical disk drive used can also have an impact on how easily the disk is read. Due to mechanical features such as the head alignment of the drive, images of the same disk captured using different disk drives can vary widely in quality (Durno, 2016; Waugh, 2014). For this reason, having multiple older disk drives on hand is useful as certain drives are better equipped at reading particular disk formats. Waugh (2014) notes that sometimes all that is needed is the persistence to make multiple attempts with the same configuration to achieve the desired results.

Once an image of the disk has been created and the files have been acquired, dealing with obsolete file formats is an additional challenge, though one that is beyond the scope of this guide (Durno, 2016; McKinley, 2015).

Best Practices

Storage conditions

As described in the previous section, proper storage of floppy disks is critical. Disks should be stored upright in clean storage cases that protect from dust and debris. Ideal storage conditions are a cool, dry place below 23°C and below 50% relative humidity and out of direct sunlight (Ahl, 1983; Dicks & Iraci, 2017; NAA, n.d.).

Description and documentation

Before beginning the digitization process, taking a photograph of the floppy disk is a recommended practice as the label often contains useful metadata about the format of the disk or its contents. Depending on space constraints, this photograph can be included in the SIP. A

physical description of the disk along with its physical condition is typically recorded in a log kept by the archive (Durno, 2016; McKinley, 2015; Sampson, 2015; Wisner, 2019).

Creating a disk image

Creating a disk image is a practice that the archives community has learned from the digital forensics community (Durno, 2016; Lee et al., 2013; Meister, 2014). A disk image is a sector by sector copy of the entire disk, including the parts of the disk not usually seen by the user. A disk image captures "existing files and file structures as well as deleted files and unallocated disk space" (Gialanella, 2018). The disk image serves a baseline copy of the artifact; any further extraction of files or manipulation of the disk contents are done on the copy (Durno, 2016, Lee et al., 2013). Disk images should be treated as the basic units of acquisition, rather than individual files. When viewing digital forensic practices through an archival lens, "using write blockers, creating full disk images and extracting data associated with files is essential to ensuring provenance, original order and chain of custody" (Lee et al., 2013, p. 18).

Best practice is to keep the disk image as well as the extracted files “to provide flexibility in future preservation actions and decision-making” (Meister, 2014, p. 12; Gialanella, 2018, UNC School of Information and Library Science, 2018). Keeping the disk image provides the possibility of providing remote, dynamic access to disk image contents in the future. Though the technology was not well distributed as of 2014, a potential future access scenario could be to give users direct access to preserved disk images that are stored on a server, in which case, users would not need to download the disk images (Lee et al., 2013, Meister, 2014).

Ensuring authenticity: Write blockers

Write blockers ensure that none of the data or metadata on the disk is altered or overwritten while the disk is being copied. This is key for ensuring the authenticity of floppy disks. Following best practices from the digital forensics community, a hardware write-blocker should always be used; the physical write block tabs on the disk are not reliable, nor should the write-block mode integrated in the software be the only source of protection (Lee et al., 2013; UNC School of Information and Library Science, 2018). Using external USB write blockers may reduce the success of the imaging process because they introduce another intermediary, therefore it is recommended to use a disk drive with integrated write blocking functionality, or a controller card with hardware write-blocking functionality such as the KryoFlux (Durno, 2016; KryoFlux Webstore, n.d.; Waugh, 2014).

Ensuring reliability: Sustainability of the disk image file format

A sustainable file format is necessary to ensure reliable long-term access to the digital content. Sustainability factors include the degree to which the format is adopted, whether the format is self-documenting (embedded metadata), and the level of disclosure, which refers to "the degree to which complete specifications and tools for validating technical integrity exist and are [openly] accessible" (Library of Congress, 2017a). The most commonly used file formats for the disk image are the Expert Witness Format (abbreviated as EWF, full name Expert Witness Compression Format, EnCase E01 Bitstream, file extension .E01), followed by the Advanced Forensic Format. Both EWF format and AFF formats store metadata, including technical metadata about the disk, any compression method used, and checksums, within the disk image file. This is useful for tracking provenance, verifying integrity, and providing some assurance that the image has not been tampered with (Durno, 2016; Lee et al., 2013). Raw stream images, however, do not provide this embedded metadata, though it may be necessary to use raw images should the disk reader be unable to read the format of the floppy disk (Durno, 2016; Garfinkel, Malan, Dubuec, Stevens, & Pham, 2006; Waugh, 2014). The AFF format, created as an open source alternative to EWF, is no longer recommended for use by its creator, as the EWF format is now openly well-documented with open source support (Durno, 2016; Library of Congress, 2017b; Garfinkel et al., 2006).

Metadata

Metadata included within the disk image is useful for tracking provenance and verifying integrity, but the true advantage of applying digital forensics techniques to digital preservation is the ability to extract metadata from the file system that is contained in the disk image (Lee et al., 2013).This is a key step in preserving the floppy disks: “Extracting the filesystem details and original structure of file objects as preservation metadata is a key step in documenting contextual details about the technical environment in which file objects were created and/or managed” (Meister, 2014, p. 13). The standard format for this metadata is the Digital Forensics XML format – known as DFXML (Lee et al., 2013). BitCurator facilitates extraction of filesystem metadata (UNC School of Information and Library Science, 2018).

Beyond the extracted forensic metadata and the metadata included within the disk image, best practice is to record the preservation activities that were carried out. Future archivists need to understand how the floppy disks and disk images were handled. Using the preservation metadata standard PREMIS, BitCurator facilitates the generation of preservation metadata which enables the archive to manage the digital object within the digital repository long term (Library of Congress, 2010; Meister, 2014; Gialanella, 2018; UNC School of Information and Library Science, 2018).

Virus scanning

Running virus scans is an important practice when working with floppy disks. Computer viruses from decades ago still pose a threat to modern computers, as floppy disks were a common attack vector (Durno, 2016). Disk images should be scanned for viruses, as the boot sector can contain viruses. In addition, the extracted files should be scanned immediately for viruses before they are handled further. ClamTK, an open source interface for the ClamAV antivirus software program, is the most commonly used tool for this purpose. It is integrated into the BitCurator environment (UNC School of Information and Library Science, 2018). In addition to running virus scans, working in a Linux operating system affords added security (Durno, 2016; McKinley, 2015).

Ensuring integrity: checksums

Throughout the digitization process of imaging, transfer, and preservation, best practice is to run checksums at multiple points to ensure that none of the data has been altered from the previous versions. MD5 and SHA-1 are two checksum algorithms that are widely used in the digital preservation community for fixity checks and are both options in BitCurator. A survey of the literature related to floppy disks found that MD5 checksums were more frequently used. A checksum should be generated as soon as the disk image is acquired to ensure the source media and the resulting image match (Durno, 2016; Gialanella, 2018; McKinley, 2015). Running checksums at other points is advised, though the exact step in the workflow when this is typically done varies. Generally, checksums are run after significant steps in the digitization process, such as when the disk image is first created, when files are extracted, and upon ingest.

Identifying personal or sensitive information

Taking advantage of forensic tools that can identify, flag and redact, or restrict access to sensitive information is a recommended practice that can save archivists a lot of time in comparison to expensive manual analysis of files. Identifying personally sensitive information is a key step that must be carried out before providing access to disk contents. Using BitCurator, a scan of the disk image and the resulting report permits the archivist to triage files that require further analysis (Lee et al., 2013, Meister, 2014; UNC School of Information and Library Science, 2018).

Best practices: Use of hardware and software

Hardware

The KryoFlux is by far the most commonly used hardware involved in preserving and digitizing floppy disks. The KryoFlux is a controller card, a hardware component that works as an interface between the motherboard and peripheral devices, which in this case is the USB disk drive (Techopedia, n.d.). The KryoFlux is known for its ability to deal with a wide range of formats, including copy protected disks, unusual formats, and Mac formatted disks. The KryoFlux is equipped to handle media that is degrading or suffering from bitrot. According to Durno (2016), it is the only device capable of imaging "flippy" disks, when used with a modified disk drive. The KryoFlux has hardware level write-blocking superior to the write block tabs on the disk themselves, eliminating the need for an additional external write blocker. (Durno, 2016; Gough, 2013; KryoFlux Webstore, n.d.; Waugh, 2014). Though KryoFlux software is also available for conducting digital preservation/forensics, BitCurator is far more commonly used. For more information about the KryoFlux, please see The Archivist's Guide to Kryoflux (Allen et al., 2018).

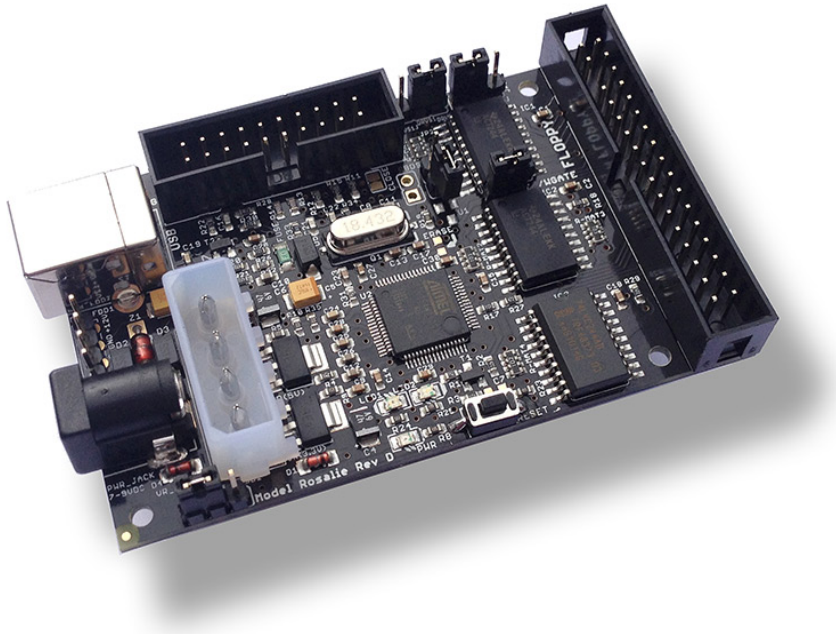

Figure 3. KryoFlux USB-based floppy controller. From KryoFlux webstore (n.d.). KryoFlux personal edition premium. Retrieved fromhttps://webstore.kryoflux.com/catalog/product_info.php?products_id=30

Software

BitCurator is one of the best-known tools for digitization and preservation of electronic media, including 3.5-inch floppy disks. Since its conception in 2016, this free, open source software has quickly become a tool of choice for digital preservation, as it was developed specifically for the LAM community. It integrates many open source tools such as ClamTK (antivirus service), TestDisk (recovery tool), write-blocker software, and Guymager (imaging and recovery service). It offers a friendly comprehensible user interface and it is easy to integrate into existing digital curation workflow. It is well documented and has excellent support forums. (Lee et al., 2013; Meister, 2014; UNC School of Information and Library Science, 2018; Waugh, 2014). The BitCurator QuickStart Guide is an excellent resource with detailed step-by-step information (UNC School of Information and Library Science, 2018).

Ethical Considerations

It is important to mention a few ethical issues regarding the digitization and preservation of floppy disks. First, when disk images of floppy disks are created using BitCurator, the deleted documents from the floppy disk are still preserved. Thus, it is possible to retrieve and access those deleted documents. It is important as an archivist to decide what should be done with these documents. According to the codes of ethics of the Association of Canadian Archivists and the International Council on Archives, archivists should ask donors about their intentions regarding deleted documents at the beginning of the acquisition process. If it is impossible to contact the donor, the code of ethics requires that the recovered deleted documents be immediately removed since the author likely did not intend to give access to these documents (International Council on Archives, 1996; Association of Canadian Archivists, 2017).

Second, it is important that there is transparency in the digitization process. In an increasingly digital world, the code of ethics of both associations requires that archivists be able to demonstrate the process behind the digitization and preservation of born-digital objects. Thus, archivists must be able to demonstrate and justify their decision-making regarding the actions they took (International Council on Archives, 1996; Association of Canadian Archivists, 2017).

Procedure for Digitization and Preservation

The following digitization and preservation workflow (see Figure 4) is based on the best practices established in the previous section and on the available resources that ARCS currently has or is in the process of acquiring. The process uses BitCurator software, which ARCS has already installed on a dedicated computer, the Targus PA905 disk drive, and the Tableau write blocker. As ARCS is in the process of acquiring the KryoFlux controller card, its functionalities within certain steps have been included in the workflow.

This workflow is adapted from Meister (2014), which divides the process into four stages: Document Media, Create Disk Image, Export Files, and Initial Analysis. However, we have added an Appraisal Stage and Pre-ingest/Ingest Stage to this workflow to account for the final actions in the preservation process. There are a total six stages which are each subdivided into steps (Figure 4).

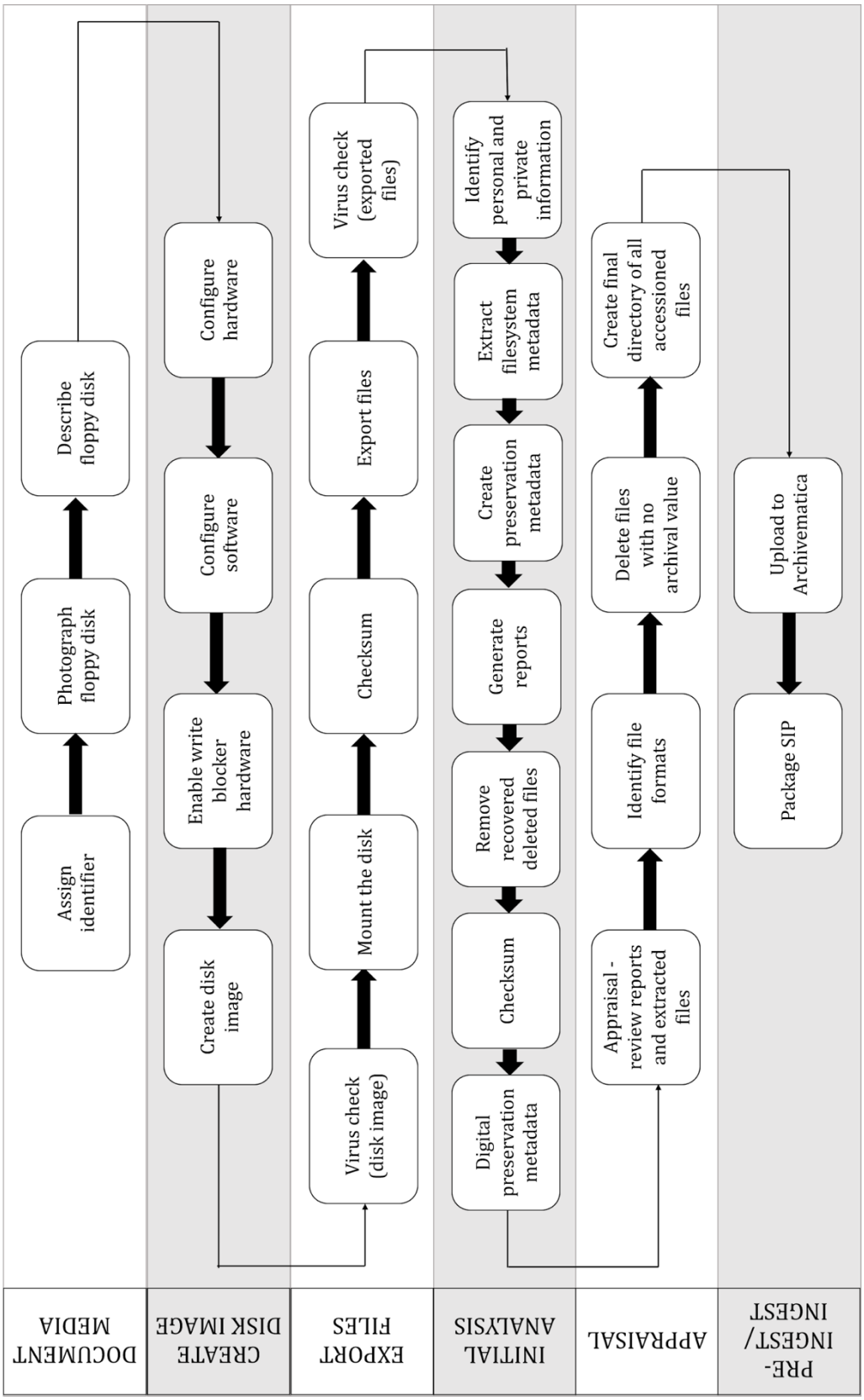

Figure 4. 3.5-inch Floppy Disks Digitization and Preservation Workflow for ARCS.

Stage 1. Document media

1. **Assign identifier**. The first step in the preservation workflow is to assign a unique identifier for each disk, if the floppies do not yet have one.
2. **Photograph floppy disk**. Next, photograph the disk.
3. **Describe floppy disk**. Document the physical characteristics of the diskette. This includes any labels, markings, or written notes present on the media and its physical condition.

These steps are important as they permit any archivist to physically identify all the floppy disks. The descriptive information collected in this step will be entered in the preservation metadata Excel sheet as well as the during the creation of the disk image. The output files expected are .jpg for the photograph and .info and .xlsx for the descriptive metadata (Meister, 2014; McKinley, 2015).

Stage 2. Create disk image

To prepare for disk imaging, the software and hardware tools need to be configured. BitCurator software, the Targus PA905 disk drive, and the Tableau write-blocker will be used.

1. **Configure hardware**. To configure the hardware, connect the disk drive and the write-blocker via USB. Turn on the write blocker and ensure it is connected to the power supply.

2. **Configure software**. To configure BitCurator, set the USB mount policy to "Read only" by clicking on the disk drive icon in the upper right corner. Create a new folder on the desktop for each 3.5-inch floppy disk being digitized following the naming convention below:

   Name of fonds_disk X

   Within the main folder, create five subfolders: Disk image, Checksums, Extracted files, Bulk extractor output, and Run all reports.

3. **Enable write blocker hardware**. In addition to using the read-only mode in BitCurator, using write-blocker hardware is strongly suggested. If the KryoFlux is available, it is preferable to use its hardware level write blocking functionality rather than the Tableau write blocker. Finally, physically switch the write-protection tab on the 3.5-inch floppy disk, in the upper right corner, to a read-only mode before inserting the floppy disk into the disk drive (see Figure 5) (Durno, 2016; Meister, 2014; Meister, 2016; UNC School of Information and Library Science, 2018).

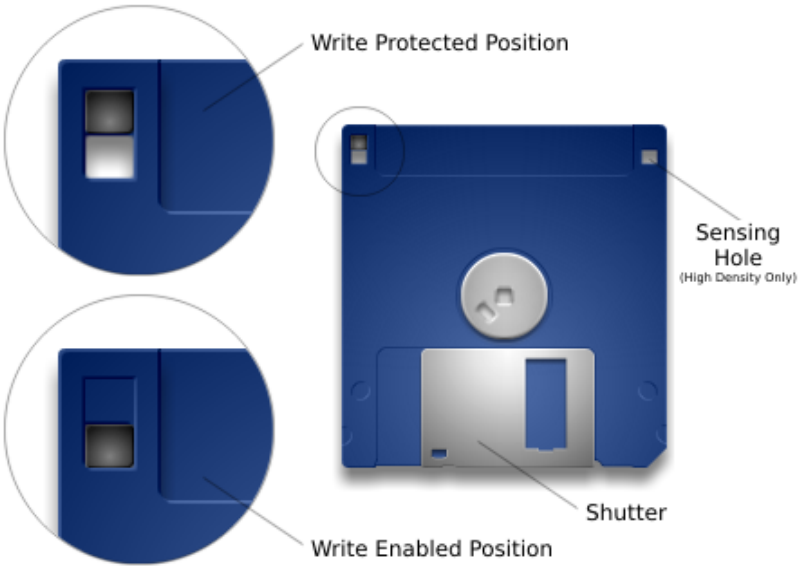

Figure 5. Write Protecting/Enabling Floppy Disks.
From WikiAmigaOS. (2014). AmigaOS Manual: Workbench Basic Operations. Retrieved from https://wiki.amigaos.net/wiki/AmigaOS_Manual:_Workbench_Basic_Operations

4. **Create disk image**. To execute this task, double-click on "Image and Recovery" desktop tool then double-click Guymager, which will prompt the entry of an administrator password. The 3.5-inch disk drive should appear as the first entry in the Guymager window; right-click on the device and select "Acquire image". In the next window, enter the following information:
    - Select "Expert Witness Format";
    - Enter the name of the fonds in "Case number";
    - Enter the disk number in "Evidence number";
    - Under "Examiner", enter the first and last initials of the person or people performing the digitization process, separating multiple people by semicolon;
    - In the description field, enter a physical description of the floppy disk, such as its color, manufacturer, information on the label, physical condition, etc.;
    - The "Notes" field is pre-populated with a serial number; leave this as is;
    - Choose the file destination by clicking the three dots; select the "Disk image" folder created in Stage 2, Step 2;
    - To access the files created on the desktop, go to Computer > Home > bcadmin > desktop > Main folder (Name of fonds_disk X) > Disk image folder;

- Enter the image filename, following the convention used for the main folder, except without spaces: Nameoffonds_diskX; and
- Leave "Calculate MD5" and "Verify image after acquisition" checked.

Click "Start". BitCurator will produce two files: the .E01 image file and a .info file containing the data produced by the disk imaging action and subsequent actions taken on the disk (Durno, 2016; Meister, 2016; UNC School of Information and Library Science, 2018). Finally, safely remove the disk by clicking on "Home" on the desktop, then clicking the eject button beside the disk drive. Remove the disk.

Stage 3. Export files

The five steps in this category are based on the digitization and preservation workflows of Purdue University Born-Digital Archives and Special Collections Content (2016) and British Library (2016):

1. **Virus check (disk image)**. Conduct a virus check on the disk image file (.E01): in the BitCurator environment, there already is a tool present, ClamTK. This antivirus tool can scan the disk image to check for any viruses. This includes a scan of the floppy disk boot sector. To execute the scan, select ClamTK under "Additional Tools," select ClamTk. Then select the "Scan a file" icon, then select the .E01 disk image file created in the previous step. A pop up will indicate if any threats were found. If there are any threats, the disk can be quarantined by selecting the "quarantine" options in BitCurator, or the infected files can be removed. If the latter is done, take note which files have been removed. To view the results of the scan, select the "History" icon and click on the current date.

2. **Mount the disk** to check if the disk is readable: right-click the disk image file and go to "Scripts > Disk Image Mount". If the floppy disk is mountable, a disk icon will appear on the desktop. If there is no disk icon, it can either mean that there was no file system, or an unrecognized file system was encountered. In the case of the latter, there are two options:
    a. Either reattempt to image the disk by repeating the previous steps beginning at Create Disk Image, ideally using a different disk drive, or
    b. Attempt to image the disk as a raw stream image. Raw stream images are difficult to work with as they are not formatted, but it makes it possible to image the disk. Imaging the disk as a raw stream would require the KryoFlux: select the stream format on the KryoFlux and select the Linux dd raw image as the imaging format rather than EWF (UNC School of Information and Library Science, 2018; Waugh, 2014).

To unmount the disk, right-click the disk icon and select “Disk Image Unmount”. A prompt for the BCadmin password will appear prior to unmounting.

3. **Checksum**. Run a checksum to ensure the integrity of the disk image: this can be done using the Nautilus tool in BitCurator. To do so, open the folder containing the .E01 disk image file, right click on the file on which a checksum should be run, then click “Scripts > File Analysis > Calculate MD5” (Meister, 2017). In the dialogue box that appears, select “Save to file”. The checksum file will automatically save in the current folder (file extensions .md5); add checksum 1 to the filename, then move the file to the checksum folder.

4. **Export files**. Use the BitCurator Disk Image Access tool found in the “Forensics and Reporting” folder. A pop-up window will automatically prompt to load a disk image. The disk image will be automatically processed, and information about the image will be presented in the top-right dock widget. Click the “Select All” icon in the toolbar, then click “Export selections”. A prompt to select a location for the files will appear; select the “Extracted Files” sub-folder that was created in step 2.2. Take note of the number of files that were exported to ensure that all files are scanned in the next step. Note that extracting a large number of files at the same time can take a long time. BitCurator also has the TestDisk tool that can recover lost sectors and analyze damaged disks (Meister, 2014; Gialanella, 2018; UNC School of Information and Library Science, 2018).

5. **Virus check (exported files)**. To run a virus check on the exported files, go to the ClamTK tool, this time selecting “Scan Directory.” Open the extracted files folder and click “Ok”. This will scan all the extracted files for threats (Gialanella, 2018; McKinley, 2015; UNC School of Information and Library Science, 2018).

Stage 4. Initial analysis

The initial analysis step involves carving the raw disk contents for features of interest, extracting filesystem metadata from the disk image to generate a DFXML listing of the file system hierarchy, matching features to files, and generating high-level reports.

1. **Identify personal and private information**. Open the Forensics and Reporting folder then double-click on the BitCurator Reporting Tool. First, click on “Launch BEViewer” to run bulk_extractor. BEViewer is the graphical front-end of the bulk_extractor tool, which operates by identifying features considered restrictive such as social security numbers, image and geolocation metadata and credit card numbers. Click on the “Tools” menu at the top of the window, and select “Run bulk_extractor”. A window will appear

where you can select which scanners to run and where to generate the report directory. Use the preset scanners established by BitCurator. In the Required Parameters at the top, click the "..." to select the Image file to scan. Click the next "..." to select the "Bulk extractor output" folder as the destination for the bulk_extractor output. Click "Submit Run". The output is a series of text files and an xml report file. The features identified by bulk_extractor can be viewed in the main Bulk Extractor Viewer window by clicking on the report name in the "Reports" section on the left.

2. Next, return to the BitCurator Reporting Tool. The Run All tab will:

    a. Run fiwalk, "a digital forensics tool that processes a disk image using the SleuthKit library and outputs its results in Digital Forensics XML" to **extract the filesystem metadata** (Kamwoods - BitCurator, 2013);

    b. Run the annotation tool to link features identified by bulk_extractor features to files; and

    c. **Generate reports** that summarize and present the data produced in all the previous tasks in a series of PDF and Excel documents, as well as **create preservation metadata**. These human-readable reports will allow the archivist to conduct an appraisal of the disk contents in the following stage.

    To Run All, BitCurator needs to know the location of the disk image, the location of the bulk extractor output generated in the previous step, and the destination for the report output. Note that the entire Bulk extractor output folder must be selected, not an individual file. Select the folder entitled "Reports run all" for the reports destination. Click "Save", then click "Run" (UNC School of Information and Library Science, 2018).

    The output files of the "Run All" action include:

    - bc_format_bargraph.pdf (file): the file formats histogram
    - bulk_extractor_report.pdf (file): high-level overview of features identified by bulk_extractor (personal and private information)
    - fiwalk_deleted_files.pdf (file): shows deleted materials found in a given partition
    - fiwalk-output.xml.xlsx (file): Excel converted DFXML output (file system metadata)
    - fiwalk_report.pdf (file): High-level overview of file system characteristics
    - format_table.pdf (file): Long-form file format names for formats shown in bar graph
    - premis.xml (file): PREMIS preservation metadata
    - features (directory): the directory of annotated features

3. **Delete recovered deleted files**. In keeping with relevant code of ethics explained in the Best Practices section of this guide, permanently delete all deleted files recovered by BitCurator. To do so, open the fiwalk_deleted_files.pdf in the Run all reports folder to view a list of all the deleted files. Then, go to the Extracted files folder to delete the files indicated in the report. Note that some deleted files may be hidden; to make these visible, click the three horizontal lines at the top right of the window, then check off "Show hidden files" from the drop down.

4. **Checksum**. Perform another checksum to ensure the integrity of the files and the disk. To perform a checksum on the extracted files, right click on the Extracted files folder, select "Scripts > File analysis > Calculate md5. An md5 file will appear upon completion; move this to checksum folder. Follow the same procedure for the disk image file. When prompted, select "Save to file". Add "checksum 2" to the file name before moving it to the checksum folder.

5. **Digital preservation metadata**. Complete the ARCS digital preservation metadata template to catalogue the digital resource created (disk image, .info file, photograph of floppy disk).

Stage 5. Appraisal

1. **Review reports and extracted files**. The reports created in the previous step enable the archivist to conduct a high-level appraisal of the disk image contents. The archivist can conduct further assessment of individual files by opening the files found in the Extracted files folder. In particular, files containing personal or private information should be examined and removed/redacted as needed. Note that the bulk_extractor_report.pdf provides an overview of how many instances there are of each feature type (e.g. telephone, email, etc.), however the individual Excel files must be opened for each feature type to see which file the feature appears in.

2. **Identify file formats**. It is likely that when conducting the appraisal, there will be files in legacy formats that cannot be opened using modern software. Appraisal decisions need to weigh the difficulty of accessing and migrating files against their potential value. The availability of special software to access files or the obsolescence of the format may impact appraisal decisions (McGuire, 2018).

   During ingest, Archivematica uses FIDO and Siegfried, both PRONOM-based programs, to identify file formats. However, best practice is to use a file identification program during the appraisal step before uploading files into Archivematica as a SIP. Standard tools such as DROID, JHOVE and FITS are not included in the BitCurator environment.

For more information about file format identification, please see the section entitled "Review and Appraisal" in the resource Handling Digital Archives Before Ingest (McGuire, 2018).

3. **Delete files with no archival value**. Following the appraisal, delete all files that have been deemed as not having archival value. Note that these files will remain part of the disk image file which will be retained in the submission information package (SIP).

4. **Create final directory of all accessioned files**. It is recommended that a final directory listing of all files being accessioned is created and included in the SIP (McGuire, 2018). Fill out the ARCs digital preservation metadata for the listing directory itself.

Stage 6. Pre-Ingest/Ingest

1. **Package SIP**. Finally, the output files generated throughout the entire process must be packaged to create a Submission Information Package (SIP). See below for more information.

2. **Upload to Archivematica**. It is recommended that a checksum be performed in Archivematica to ensure the integrity of the files and the disk image. Archivematica will generate a checksum upon transfer; it is also possible for Archivematica to include pre-existing checksums generated in BitCurator in its verification (Archivematica, n.d.). Note that it will likely be necessary for files to undergo normalization in Archivematica (McGuire, 2018).

Once the digitization process is done, the archivist must address the question of whether to keep the original 3.5-inch floppy disks for long term preservation. It is hard to tell what to do exactly with a diskette after it has been digitized due to the lack of research on the subject. After extensively researching this subject, it was impossible to find any literature that mentioned what was done with the physical diskette once a Submission Information Package and an Archival Information Package was created for the disk image and extracted files. Due to the lack of research, it is only possible for us to propose a solution based on our own experience. Since the disk image is essentially a digital replica of the physical floppy disk, including all parts of the disk that are usually not visible to users, it may be sufficient to retain the disk image for long-term preservation. Preserving the disk image is akin to preserving the physical disk, but in a format that can be accessed and manipulated in the future should the physical disk become unreadable.

Submission Information Package (SIP)

Based on the available literature, the best practice for creating a SIP is to include the output files created from each step of the digitization and preservation workflow process. Literature on the topic recommends that the following output files are included in the SIP (Durno, 2016; Gialanella, 2018; McKinley, 2015; Meister, 2014; Sampson, 2015; The University of Maryland Libraries, 2016):

- photographs of the floppy disk,
- the disk image,
- extracted files,
- file directory,
- checksums,
- output from virus scans,
- generated reports,
- extracted file system metadata,
- preservation metadata,
- final file directory listing, and
- descriptive metadata about the disk.

Table 1 includes all the steps present in the digitization and preservation workflow and whether they will produce files to be included in the SIP. If yes, it also indicates, where possible, the format of the files.

Table 1. Output files and their file formats.

Categories	Steps	Output Files	File Formats
Document media	Assign identifier	Yes	contained in .info file
	Photograph floppy disk	Yes	.jpg file
	Describe floppy disk	Yes	contained in .info file and .xlsx file
Create disk image	Configure hardware	No	n/a
	Configure software	No	n/a
	Enable write blocker hardware	No	n/a
	Create disk image	Yes	.E01 image file and .info file

Categories	Steps	Output Files	File Formats
Export files	Virus check disk image	Yes	.ficlam file
	Mount the disk	No	n/a
	Checksum	Yes	.md5 file
	Export files	Yes	Various formats (depending on contents of disk
	Virus check exported files	Yes	.ficlam file
Initial Analysis	Identify personal and private information	Yes	.txt files
	Extract filesystem metadata	Yes	.xml file (DFXML format)
	Create preservation metadata	Yes	premis.xml file
	Generate reports	Yes	.pdf files and .xml files; see the Initial Analysis step for a complete list
	Remove recovered deleted files	No	n/a
	Checksum	Yes	.md5 file
	Digital preservation metadata	Yes	.xlsx file
Appraisal	Review reports and extracted files	No	n/a
	Identify file formats	Yes	TBD by archivists
	Delete files with no archival value	No	n/a
	Create final directory of all accessioned files	Yes	Text file such as .txt or .docx

To create the SIP, Sampson (2015) and Gialanella (2018) both used the Bagger software application to create a BagIt bag. A BagIt bag is a group of digital files which conforms to the BagIt standard, which specifies the folder structure of the bag (World Digital Library, 2012). The output files listed in Table 1 will be packaged into a bag, then uploaded into Archivematica.

Omeka Blog

This guide focuses on the final version of the digitization and preservation workflow established with ARCS. To learn more about the challenges we encountered while doing our test runs, please see the Omeka blog exhibit: https://biblio.uottawa.ca/omeka2/linking-cultures/exhibits/show/diskette. The blog concentrates on the process undertaken to create the final version of the workflow. It explains the challenges encountered during testing and the solutions to these problems. The blog also includes a section dedicated to information about the fonds that the 3.5-inch floppy disks used for this project came from.

References

Allen, J., Arroyo-Ramirez, E., Bolding, K., Charlton, F., Ciccone, P., Eadon, Y., Farrell, M., Hughes, A., Maches, V., Peltzman, S., Prael, A., Reed, S., Waugh, D. (2018). The archivist's guide to KryoFlux. Retrieved from https://github.com/archivistsguidetokryoflux/archivists-guide-to-kryoflux

Ahl, D. (1983). Floppy disk handling and storage. C*reative computing, 9*(12), 205. Retrieved from https://www.atarimagazines.com/creative/v9n12/205_Floppy_disk_handling_and_.php

Andolsen, A. (2006). Will your records be there when you need them? *Information Management Journal*, *40*(3), 56–61.

Archivematica. (n.d.). FAQ. Retrieved from https://www.archivematica.org/en/docs/archivematica-1.6/getting-started/troubleshooting/faq/

Archive Team. (2018). Rescuing floppy disks. Retrieved from https://www.archiveteam.org/index.php?title=Rescuing_Floppy_Disks

Archives and Special Collections. (2018). Monique Frize fonds. Retrieved from https://biblio.uottawa.ca/atom/index.php/monique-frize-fonds

Association of Canadian Archivists. (2017). Code of Ethics and Professional Conduct. Retrieve from https://archivists.ca/resources/Documents/Governance%20and%20Structure/20171018_aca_code_of_ethics_final.pdf

British Library. (2016). "Flashback" workflow for legacy content extraction. Retrieved from https://bitcuratorconsortium.org/workflows/british-library-workflow

Brown, G. (2001). How floppy drives work. Retrieved from https://computer.howstuffworks.com/floppy-disk-drive1.htm

Dicks, C. & Iraci, J. (2017). Computer hard disks and diskettes – FAQ. Retrieved from https://www.canada.ca/en/conservation-institute/services/care-objects/electronic-media/computer-hard-disks-diskettes-faq.html#faq1

Disha Experts. (2019). *Guide to RRB Junior Engineer Stage II Mechanical & Allied Engineering 3rd Edition*. New Delhi: Disha Publications.

Durno, J. (2016). Digital archaeology and/or forensics: Working with floppy disks from the 1980s. *Code4Lib Journal*, (34), 1–1.

Garfinkel, S., Malan, D., Dubec, K., Stevens, C., & Pham., C. (2006). Advanced forensic format: an open extensible format for disk imaging. Second Annual IFIP WG 11.9 International

Conference on Digital Forensics (pp 17-31). Orlando, Florida. Retrieved from https://cs.harvard.edu/malan/publications/aff.pdf

Gialanella, L.A. (2018). Disk Imaging for Preservation: Part 1. Retrieved from https://www.lib.umich.edu/blogs/bits-and-pieces/disk-imaging-preservation-part-1

Gough, L. (2013). Project KryoFlux—part 2: why bother with it? Retrieved from https://goughlui.com/2013/04/21/project-kryoflux-part-2-why-bother-with-it/

Indiana University. (2018). What are boot sector viruses, and how can I prevent them? Retrieved from https://kb.iu.edu/d/ahll

International Business Machines Corporation. (n.d.). The floppy disk. Retrieved from https://www.ibm.com/ibm/history/ibm100/us/en/icons/floppy/

International Council on Archives. (1996). Code of Ethics. Tiré de https://www.ica.org/sites/default/files/ICA_1996-09-06_code%20of%20ethics_EN.pdf\

Kamwoods - BitCurator. (2013). BitCurator Webinar on bulk_extractor, fiwalk, and the BitCurator Reporting Tool. Retrieved from https://bitcurator.net/2013/12/11/bitcurator-webinar-on-bulk_extractor-fiwalk-and-the-bitcurator-reporting-tool/

KryoFlux Webstore (n.d.). KryoFlux personal edition premium. Retrieved from https://webstore.kryoflux.com/catalog/product_info.php?products_id=30

Lee, C., Woods, K., Kirschenbaum, M., & Chassanoff, A. (2013). From bitstreams to heritage: Putting digital forensics into practice in collecting institutions. Retrieved from https://drum.lib.umd.edu/bitstream/handle/1903/14736/bitstreams-to-heritage.pdf?sequence=1&isAllowed=y

LeFurgy, B. (2012, April 11). Floppy disks are dead, long live floppy disks. Retrieved from https://blogs.loc.gov/thesignal/2012/04/floppy-disks-are-dead-long-live-floppy-disks/

Library of Congress (2017b). Expert Witness Disk Image, EnCase E01 Bitstream. Retrieved from https://www.loc.gov/preservation/digital/formats/fdd/fdd000408.shtml

Library of Congress. (2010). PREMIS for digital preservation. Retrieved from http://www.digitalpreservation.gov/series/challenge/premis.html

Library of Congress. (2017a). Sustainability factors. Retrieved from https://www.loc.gov/preservation/digital/formats/sustain/sustain.shtml

McGuire, J. (2018). Handling Digital Archives Before Ingest. Retrieved from https://learn.scholarsportal.info/all-guides/handling-digital-archives/

McKinley, M. (2015). Imaging digital media for preservation with LAMMP. *The Electronic Media Review, 3*, 89-96.

Meister, S. (2016). Creating a disk image using guymager. Retrieved from https://confluence.educopia.org/display/BC/Creating+a+Disk+Image+Using+Guymager

Meister, S. (2017). Calculate and display MD5 sums. Retrieved from https://confluence.educopia.org/display/BC/Calculate+and+Display+MD5+Sums

Meister, S., & Chassanoff, A. (2014). Integrating digital forensics techniques into curatorial tasks: A case study. *International Journal of Digital Curation*, *9*, 6–16. https://doi.org/10.2218/ijdc.v9i2.325

National Archives of Australia (n.d.). Frequently asked questions - preserving physical records. Retrieved from https://webarchive.nla.gov.au/wayback/20110314201706/http:/www.naa.gov.au/ records-management/secure-and-store/physical-preservation/faq/magnetic-tape.aspx

Purdue University. (2016). Purdue university workflow map for born-digital archives and special collections. Retrieved from https://bitcuratorconsortium.org/workflows/purdue-university-workflow

Salter, J. (2014, January 15). Bitrot and atomic COWs: Inside "next-gen" filesystems. Retrieved from https://web.archive.org/web/20150306225935/http://arstechnica.com/information-technology/2014/01/bitrot-and-atomic-cows-inside-next-gen-filesystems/

Sampson, W. (2015). Guest post: Walker Sampson on disk imaging workflow. Retrieved from https://bitcurator.net/category/use-cases/

The Science Explorer. (n.d.). Dissect a disk – find out what's inside a floppy disk. Retrieved from https://www.exploratorium.edu/science_explorer/dissect_disk.html

The University of Maryland Libraries. (2016). Processing workflow. Retrieved from https://bitcuratorconsortium.org/workflows/processing-workflow

University of North Carolina (UNC) School of Information and Library Science. (2018). BitCurator Quick Start Guide. Retrieved from https://distro.ibiblio.org/bitcurator/docs/BitCurator-Quickstart-v2.pdf

Vaughan-Nichols, S. (2017). The history of the floppy disk. Retrieved from https://www.hpe.com/us/en/insights/articles/the-history-of-the-floppy-disk-1703.html

Waugh, D. (2014). A dogged pursuit: capturing forensic images of 3.5" floppy disks. P*ractical Technology for Archives, (2)*. Retrieved from https://practicaltechnologyforarchives.org/issue2_waugh/#ftn4

WikiAmigaOS. (2014). AmigaOS Manual: workbench basic operations. Retrieved from https://wiki.amigaos.net/wiki/AmigaOS_Manual:_Workbench_Basic_Operations

Wikipedia contributors. (2019, September 17). Data degradation. In *Wikipedia, The Free Encyclopedia*. Retrieved on October 14, 2019, from https://en.wikipedia.org/w/index.php?title=Data_degradation&oldid=916261885

Wisner, M. (2019). 3.5" floppy disks: BitCurator workflow. Retrieved from https://wiki.harvard.edu/confluence/pages/viewpage.action?pageId=220660945

World Digital Library (2012). Bagger overview. Retrieved from https://project.wdl.org/arab_peninsula/workshop2012/en/doha_workshop_2012_bagger_en.pdf