A Case Study on Retrieval of Data from 8-inch Disks

Of the Importance of Hardware Repositories for Digital Preservation

Denise de Vries
Flinders University
GPO Box 2100
Adelaide, South Australia 5001
denise.devries@flinders.edu.au

Dirk von Suchodoletz
University of Freiburg
Hermann-Herder Str. 10
79104 Freiburg, Germany
dirk.von.suchodoletz@rz.uni-freiburg.de

Willibald Meyer
University of Freiburg
Hermann-Herder Str. 10
79104 Freiburg, Germany
willibald.meyer@rz.uni-freiburg.de

ABSTRACT

There are still a significant number of born-digital artefacts which have yet to be received by memory institutions. These include the works of famous or important authors, artists, politicians, and musicians, some of which are stored on digital media dating back 30 years or more. Many digital objects have escaped the blight of bit rot, and legacy hardware, though rare, is still available for reading the media. Assembling the required equipment and expertise is possible but increasingly difficult.

This is an ongoing challenge which is not confined to floppy disks or obsolete hard drives. The challenge is exacerbated by the reality that many artefacts are not considered for preservation until the creator has passed away. In addition to the past 30 years' worth of media, today's storage media will also soon be superseded with newer technology and form a new challenge for future archivists. In order to preserve the content, or the look and feel of the original objects, precautionary actions must be taken which require both knowledge and the right equipment.

Creating disk images from obsolete magnetic media is not a simple task, especially when it concerns more obscure or older formats such as 8-inch disks. There are many challenges to overcome, with each small success leading to a new phase of discovery.

In this case study we recount our progress, through many stages, of dealing with the described challenges with regards to a set of 8-inch magnetic floppy disk media. At the beginning of this project, most requirements for recovering the data from this media were unknown: the original hardware system, the original operating system, and the disk format. In this paper, we describe our approach to uncovering this information, leading to a successful preliminary outcome. This is a cautionary tale which aims to provide some lessons for use in related contexts.

CCS CONCEPTS

•Information systems →Magnetic disks;

KEYWORDS

Digital preservation; disk imaging; disk format; digital heritage

1 INTRODUCTION

There is very little literature in the digital preservation discipline regarding the challenges that must be overcome to retrieve data from obsolete media and render it accessible, complete and accurate. For instance [20] has no use cases on 8-inch disks and the actual information referenced is often rather scarce [4]. Many practitioners in digital libraries and archives presume that someone is aware of the technological changes and takes the necessary action to mitigate it. But digital preservation is still a relatively young discipline and many IT professionals have still not placed sufficient priority on media migration. Additionally, well-established workflows to hand over materials from governance offices to archives still use 20–30 year pre-transfer retention periods, which were reasonable for traditional paper-based records but are much too long for any digital storage medium to be held without intervention by media preservation experts. The situation is even worse in the personal sphere: written documents of famous authors or politicians are still valuable after their death [9] but there is little to no availability of standard processes or policies regarding the transfer of any digital equipment and/or storage media at this difficult moment. While organizational knowledge exists that *There is little existent literature that would recommend removal media of any type to be a worthy archival medium.* [10], this is not necessarily the case for the personal domain.

This case study presents the challenges and difficulties faced when attempting to save research data from 8-inch disks and the digital forensic methods needed to succeed. From the very start of this endeavour we had very little information from which to organize our work.

The brief given to us was succinct. *There is a set of 8-inch disks from early- to mid-1980s which contain research data. These data are required for current research. Retrieve the files from the disks and make them available on current media.*

The hand-over procedure did not generate much information, seventeen 8-inch disks with some additional paper stuck in a few of the disk sleeves were passed to us (Figure 1). The floppy labels of the various brands (e.g. BASF and HP) did not provide any information on the hardware or systems used to create them or on the formatting used.

The few sheets of papers accompanying the disks seemed to indicate that no file system was used at all. It appeared to be head, sector and track information. While the column containing the head information (0 for first side, 1 for the second side of a two sided medium) was rather obvious to deduce, it took a few comparisons to understand the meaning of the sector and track information. There is no record within the university about the hardware systems used at the time and inquiries within the organization only hinted that they had introduced office writers (electronic IBM typewriters) and later CPM machines, both featuring 8-inch drives; Unfortunately, nobody had a clear memory of the exact systems. Due to the number of Hewlett-Packard (HP) disks in the set, it was presumed that an

Figure 1: Sample of the disks by different manufacturers evaluated in the study

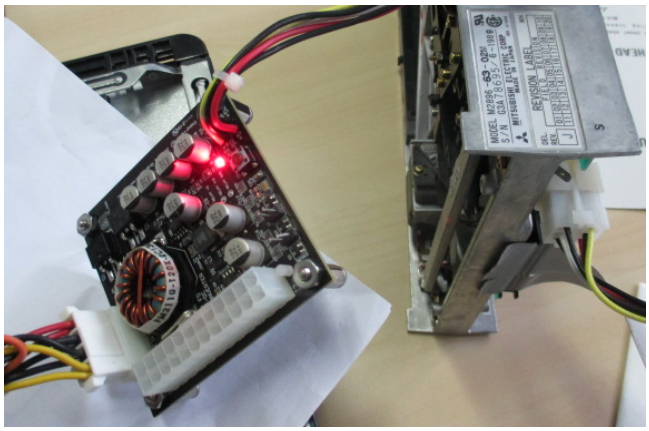

Figure 2: Providing the 8-inch drive with 5V and 24V

HP system could have been used and further investigation into the part numbers was tried.

The sheets, with some of the disks, list information formatted as

```
NAME PRO    TYPE    REC/FILE    BYTES/REC    ADDRESS
H8,0,1 148
ALT-A           DATA     1          1188       0/1/0
```

This we determined was from a HP CAT command, which is interpreted as probably from an HP 984535 series. H is the drive type, 8 the controller select code, 0 the drive address, and 1 the unit address. The first entry, in this example, indicates a large data file of 1188 bytes, file type unknown.

Extracting data from 5.25-inch and 3.5-inch floppy disks, which were used more recently and more commonly, is readily achievable as it is still possible to find systems which feature a 3.5-inch drive and systems which have at least a floppy controller present. At a computer centre it is still often possible to find all components necessary in order to get working hardware set up to read DOS formatted disks (MFM recording format, [12]), which was the most common format.

The case for 8-inch disks is more complicated. Eight-inch floppy disks were generally out of use by the beginning/mid-1990s and were not a common part of typical office or personal computer systems at that time; thus, they never obtained the interest of a wider enthusiast scene as did the home computers of the same era. The 8-inch disks have a completely different form factor and different connectors for both data and power supply. Thus, there is no standard support for them in most operating systems, and very little general information on the Internet.

2 HARDWARE REQUIREMENTS

The status of the research data was not entirely clear and there was some indication that the disks might contain private (clinical) information on living persons. Thus, the disks could not be sent off-site. As no service company or expert who was knowledgeable about these disks could be found within the vicinity, we started our study with the selection and acquisition of the hardware required to capture the bits.

The initial step was to find an appropriate machine which could both connect an 8-inch floppy drive and host a recent operating system to handle reading the data and transfer it to current media.

Two machines were evaluated: An older HP AMD CPU office desktop machine (from approximately 2008) and a more recent DELL Optiplex system, both still featuring a floppy controller. The BIOS of the HP allows setting a single floppy to either 3.5-inch 1.44 Mbyte or 5.25-inch 1.2 Mbyte. The DELL BIOS made it more difficult to configure for the older floppies, as only the 3.5-inch option could be set. Additionally to make things more challenging, actually attaching a 5.25-inch or 8-inch drive to the DELL machine electrically required a power cable adaptor from the more recent 3.5-inch SATA/HDD connectors to the old peripheral power connector style 4 pin.

As no 8-inch disk drive was to be found in the university, a "new" one had to be bought. The market, in the 1980s, offered a wide range of drives each with their own special features. A half-height drive not requiring AC power was chosen. A Mitsubishi (M2896-63) 8-inch drive in mint condition was acquired for about 240 € from eBay. In addition, a converter for the power supply of the 5/24V of the drive and a floppy disk adapter for 8-inch drives on PC floppy controllers were procured from DBit [5] for a cost only marginally less than the drive. The necessary cables were scavenged from old systems and recycling bins (e.g. a 50-pin ribbon SCSI cable and multiple standard floppy attachment cables) and refitted to be used with the 34- and 50-pin connector of the drive [6] (Figure 3). The power converter either plugged into a spare molex power connector from the PC's power supply or could be powered from a standard ATX mainboard power input (Figure 2).

Both the 24V power supply and the cable adapter board have power requirements and fit traditional PC power cable sockets. To connect the 8-inch drive without external power supplies, the machine needed to have a 3.5-inch floppy disk power connector and a standard molex power connector (for the 5.25-inch and 8-inch drives, the latter one via the voltage adaptor). Having wired all this, the setup for the 8-inch floppy reading experiment reached an initial working status.

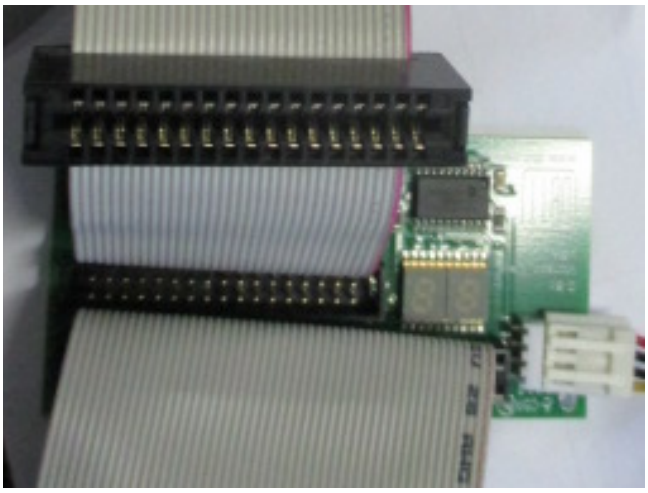

Figure 3: Electrically adapting the drive and providing a track readout

Checking the system

As no-one was really familiar with the drive, some preliminary tests were run with standard floppy drives. To create a baseline for the experiments, the two machines were checked with both 3.5-inch and 5.25-inch drives and subsequent floppy reads (1.44 MB floppies for 3.5-inch drive and both 360 KB and 1.2 MB floppies in the 5.25-inch drive). We imaged 5.25-inch DOS formatted floppy disks using the Linux `dd_rescue` command and Linux `mtools`. The drives and controllers were confirmed to be working properly.

The BIOS settings in the PC did not need to change as they can be overridden when loading the Linux kernel. The only prerequisite is an enabled floppy controller. The Linux module `floppy.ko` takes (repeated) parameters in the form of `floppy=0,1` (indicating the physical floppy drive and the type of floppy respectively). Nevertheless, the Linux kernel is not aware of 8-inch drives and has no specific settings for them in the module.

The 8-inch drive was connected and the BIOS of the machine configured to 5.25-inch drive. It could be seen, by the flashing access LED, that the drive reacted to commands issued by the controller, for example, when loading the kernel module or running a program which required access to the drive.

3 PRELIMINARY READING AND WRITING TESTS

To conduct some preliminary experiments on the drive, an unused 8-inch floppy (manufactured by Nashua), of unknown quality, was taken from a display case of obsolete media in the computer centre. This was done to run some reading and writing experiments to see if the floppy could be formatted and read without endangering the original disks. One of the challenges was the unknown characteristics of both the floppies and the drive. For the latter there was a specification sheet which came with the drive. It was not apparent to which degree the controller/drive combination would be able to read floppies, which were created on completely different system(s). The same applied to the differences of the kernel configuration which only officially knows about 5.25-inch drives and has no knowledge about 8-inch drives. Unfortunately, our research revealed that there were multiple systems which used 8-inch drives quite differently with regard to format, capacity and geometry (e.g. just for IBM floppies of that format, [26]).

When doing these first trials, the kernel or tool messages indicated an "active write protection". The re-configuration of the drive to WP (honour write protect) or NP (not protected) did not change anything in the ongoing formatting tests with `fdformat`. Mysteriously, after a while the "active write protection" disappeared for an unknown reason.

It was possible to low-level format the test Nashua floppy disk using a rather wide range of settings, but the verification failed for every setting. It is known that these formatting tests run smoothly on 5.25-inch floppy drives by a stepwise movement of the read/write head. However, on the 8-inch drive some settings, especially the higher density ones, produced strange jumps of the head and odd sounds. The DBit controller displayed the same track number on its two-digit seven-segment display as the `fdformat` gave for its operation (Figure 3).

Finally we did a simple reading test by inserting the disk and executing a directory listing command (`ls /dev/fd0` and `mdir a:`). All "read" tests of different disks, including those from the archive, failed as track 0 could not be read. This was a first indication that the disks were not of DOS environments origin, possibly complicating the access significantly.

4 USING THE BITCURATOR LINUX ENVIRONMENT

To rule out hardware-related issues such as cabling, the first suboptimal connection – a long cable with an open end which might lead to signal reflections – was shortened and a second edge connector cable was produced. The new one avoided any malfunction of the connector as it was an original 50pin.

A PC was then booted from a USB stick containing the BitCurator suite to avoid any hardware emulation a virtualization layer might introduce. The number of tools in this particular setup was limited to the command line floppy and raw IO tools available with the Linux operating system, such as: `getfdprm`, `setfdprm`, `fdrawcmd`, `fdformat`, `mtools`, `dd`, and `dd_rescue`.

One of the major challenges was the lack of experience with the 8-inch drives and floppy disks. Thus, trial and error was the approach used. The knowledge needed to be unearthed step-by-step hoping to find the relevant bits and pieces to get everything working in the intended way.

For this task the web site of the Linux floppy disk utilities [25] was consulted as the Linux man (manual) page only explains the command line settings but does not give examples. We reproduced various suggestions provided from the "How to identify an unknown disk using Fdutils" section. "Finding out the number of sides (heads)" helped us to determine that the test disk we used was actually single sided and that two sided formatting and verifying must fail. Further configuration parameters were checked and tested using `fdrawcmd` and `setfdprm`, followed by a successful low-level formatting tool. Error messages were checked from both

the tool's printout and the kernel log file. The need within the wider Linux community for floppy disk tools will diminish further and the web site, last updated in May 2005, is in risk of disappearing at some point leaving no or only a few traces.

Disk Geometry and other Parameters

For non-familiar users the options for the disk geometry remain a bit opaque, apart from the recording mechanism which is presumed to be MFM and cannot be changed. For example, typically the drive supports 77 cylinders (should equal number of tracks), but the IBM specification for type 2D (printed on the BASF disks) have a special index cylinder and then 74 usable cylinders [26]. While the index cylinder configuration remains the same, the remainder can differ significantly.

The label on the Nashua test disk revealed very little. There is a part number (FD-1D-WP) which we were able to verify is a single-sided double-density floppy disk. The specifications on the original manufacturer's box listed on the side of the carton are as follows: FD-1D WP-R (the item number), S.S. / D.D. SOFT which translates to single sided, double density, and soft sectors. This explained the failures of the previous low-level formatting verification experiment: A single-sided disk can actually be formatted as double sided, but reading from the empty side will fail. Changing the parameters to single sided with `setfdmprm` progressed our attempts at verification.

The `fdrawcmd` can issue raw floppy controller commands in order to discover the data transfer rate, the sector numbering scheme, and the number of sides on a disk. We successfully ran this on the test disk to confirm the geometry. Unfortunately, we were not successful when trying the target disks. We had to rule out possible problems such as:

- The disks were not formatted at all. Even though these were backup disks, as implied by the labels, that is no guarantee that they had ever been verified.
- The disks may have been improperly stored before they were transferred into the university archive, thus losing most of the magnetic charge.
- The drive simply cannot interpret the recording format.

5 USING AN OSCILLOSCOPE

To identify if the disks contained any recorded information at all, an oscilloscope was connected to the analogue amplifier behind the magnetic reading head. A GW Instek GDS-1022 oscilloscope was linked to the disk drive at AD3 and AD4 on the read circuit (Figure 4). The signal was set to normal level 2.8V on TP-AD3 and 2.75V on TP-AD4. Signal peaks for both channels were 800Vpp.

The rotational speed of the 8-inch floppy drive is rather slow compared to modern equipment and the recording density is very low. Even with just 26 sectors of 128 Bytes, a substantial number of changes in the magnetic flux are required to represent the data. It was hoped that it would be possible to at least distinguish between an unformatted disk (or demagnetized disk due to bit rot) and a disk containing data: white noise in the first case and some visible, repeated changes in the second.

To trigger readings, the low-level program `fdrawcmd` was run with different options on several disks [25]. Initially, we focused

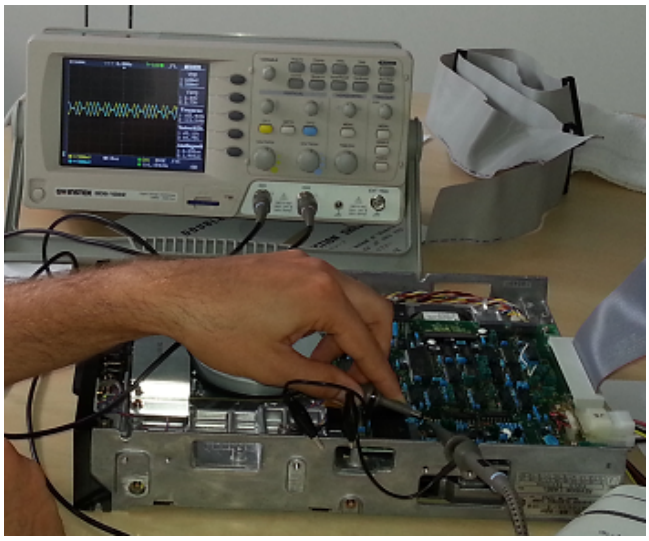

Figure 4: Connecting an oscilloscope to verify magnetized structures and check for patterns

on track 0 and started with the Nashua disk, using `fdrawcmd` to produce read signals, which produced a visible pattern of different frequencies on the oscilloscope. In further rounds, we looked for similar patterns on track 0 of the other disks. The picture obtained was quite different and looked more like white noise compared to the measurements on the Nashua.

Changing to a higher track number produced, for both classes of disks, patterns which were definitely visible. This could explain why we were not able to read anything with the methods we had been applying up to that point.

The oscilloscope setup had some limitations for the purpose of data interpretation. We were unable to store longer sequences of signal readings but could only freeze the content of the screen (Figure 5). For further stream interpretation a logic analyzer would be useful, e.g. a Bitscope Micro with DSO Data Recorder software [1].

The readings from the disks did not imply that they were completely empty. We had confirmed that our hardware setup of the 8-inch drive and the adapters were working, and that the disks seemed to contain something, our next step was to employ some more advanced approaches, such as forensic floppy controllers.

6 KRYOFLUX AND CATWEASEL

Having established that the delivered disks were not empty, different approaches were chosen. These circumvented the use of the standard floppy controller to allow a more direct access to the medium [13]. The KryoFlux and Catweasel floppy controllers are add-in cards which were designed to read disks at a low-level approach and circumvent most of the built-in PC floppy disk controller logic.

As the University of Freiburg does not have either of these devices, we asked for help from our colleagues at the Computerspielemuseum (Computer Games Museum), Berlin who have both

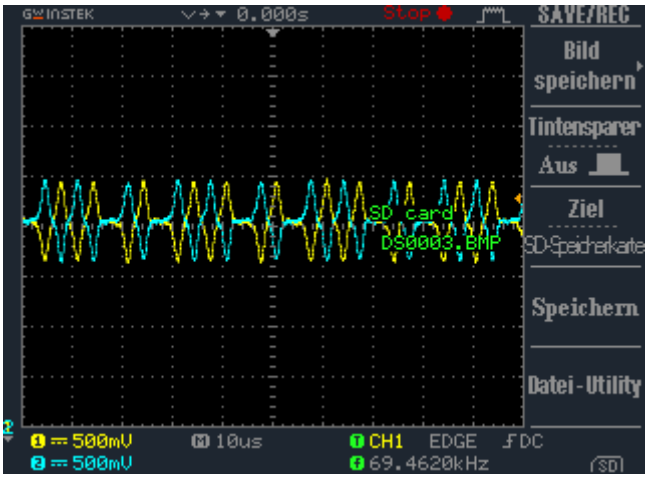

Figure 5: Oscilloscope screen capture

a KryoFlux and a Catweasel for their huge collection of magnetic media from a wide range of computer and gaming systems.

The KryoFlux

The KryoFlux card is built into a tower case containing a power supply unit (PSU), a 5.25-inch drive and a 3.5-inch drive. The Mitsubishi drive was connected to the KryoFlux which in turn was connected to a Dell laptop running Windows XP. The 50-pin data cable was connected to the fdadap floppy disk adapter, then the 34-pin to the KryoFlux. The KryoFlux was connected via USB cable to the laptop. The 5V/24V power supply to the disk drive was provided as previously described.

For our initial test of our setup we used our test disk, which failed after reading only the first track. As this test disk has consistently failed, we concluded that it is a dud disk. We then tried to calibrate the drive. This was a bad move, as it calibrated only to track 0 and subsequent attempts at imaging also failed.

The Catweasel

Please note that the Catweasel is no longer manufactured as the demand and interest in magnetic floppy media has diminished. We knew of software that worked with Catweasel to image 8-inch disks, so we connected the Catweasel system to our disk drive and tried to run the Catweasel software `ImageTools 3`. However, as there is no configuration for 8-inch disks this did not work out. We then tried `cw2dmk` [17], as the documentation mentions 8-inch disks and that a variety of formats can be imaged. Unfortunately, this also did not work, returning the error "Failed to detect any drives". To make sure that the Catweasel device was in good working order, the 5.25-inch drive was reconnected and we successfully imaged a Commodore 64 disk. Retrying the 8-inch drive, `cw2dmk` failed again with the same error. Later on it was discovered that `cw2dmk` does not work with Windows XP. Another backwards step!

The KryoFlux Again

After connecting it all up, we powered up according to the KryoFlux manual but the software reported that it did not recognize hardware. This was strange as we did have it working earlier in the day.

Checking everything again, we noticed that the circuitry was getting power from two sources and deduced that this was preventing the KryoFlux from resetting. We then unplugged power to all components except the laptop, rebooted and then plugged in the KryoFlux and then powered on all downstream components. This was the correct way to do it. The KryoFlux recognized the drive and we were able to start imaging.

The imaging was set to output KryoFlux Stream Files, preservation and all other settings to the default values. This meant that 83 tracks were tried on each side of the disk, resulting in "Error reading stream device" from track 77 onwards (with tracks starting at 0).

We managed to get an image for each disk, albeit in so far uninterpreted stream files. The stream data contain two items of logical data, the timing of flux transitions and timing of the index. This is just a very low-level interpretation which does not translate at all into human-readable form without further transformation. Each disk took around 4 to 5 minutes to image.

7 ANALYZING THE MAGNETIC FLUX

Diskettes are magnetic storage media on which small areas are more weakly or more strongly magnetized. The magnetic read heads on a disk drive operate by detecting the changes in the magnetic flux. The *encoding* specifies how the bit is written to the disk surface, and the disk *format* specifies the byte sequences to represent the structure of the data, such as the cylinder, head number and sector number.

A disk format is the organization of data on the disk which enables the computer system to recognise and verify the data. The bitstream (flux) is separated into addressable parts – sectors – with data marks and address marks in order to tell different types of information within the sector apart and to perform error checking. The binary information is encoded as a pattern of magnetic flux reversals, that is, changes from "0" to "1".

Recordings are made on a single track at a time. A track consists of a serial sequence of bits which are all interpreted as 8-bit bytes. To ensure that there are changes in magnetisation, an *encoding* method is used.

The read electronics must be able to get and maintain bit and byte synchronization. Bit synchronization is achieved with a clock bit, i.e., encodings have data bits separated by clock bits.

In MFM encoding, the first pulse period always contains a clock pulse; the second pulse period may or may not contain a data pulse. If the digital data is a "1", a data pulse will be present in the second pulse period. But, if the digital data is a "0" then there is no pulse present (see Figure 6).

The KryoFlux stream format [16] is recorded in binary, and records only the changes in the magnetic flux of a given track with timing information and cannot reveal any bits directly. The KryoFlux software suite contains stream analyzers for a wide range of different recording encodings for a variety of systems, e.g. FM, MFM and so forth. It was unclear on which system the floppies

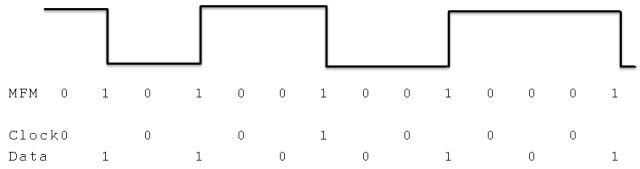

Figure 6: MFM encoding

were actually written, so all options for different encodings were tried and all failed.

In the meantime, we had identified the drive type as an HP9895A and researched the drive specifications from the resources at the HP Computer Museum [3] and The HP 9845 Project [14]. It was discovered that this disk drive could read and write both Frequency Modulated (FM) and Modified Modified Frequency Modulated (MMFM or M2FM [12]) encodings. The KryoFlux software had not recognized an FM encoding, so we surmised that the encoding was possibly M2FM, with Least Byte First and Least Bit First. While these sites have a wealth of information, documentation of the M2FM format is missing.

Creating Logic Analyzer Software

A small Python script was written to reproduce the recording patterns to be analyzed and visualized in logic analyzer software. A major challenge was to derive the clock signal from the recording as all possible encodings are self-clocked (Figure 6). While the low-level formatting creates quite clean recording patterns the actual data written to sectors often implies tiny distortions (jitter) as it is impossible for the mechanics of the drive to perfectly sync to the clock. This step was really helpful as we were now able to scan along the stream of each track. Our first interpretations were:

- The recording format appeared to be FM (and not M2FM or MFM as we expected to see).
- The low-level format of the system produced clearly visible synchronization patterns at regular distances. One followed by a short block of data, the next one followed by a long block. These patterns were visible for every track.
- This was interpreted as possibly a kind of an information header to the sector that follows, which presumably contains the actual data.
- Counting the zeros – which are easier to count, especially in an empty sector – in the graphical representation we estimated a sector size of 128 Bytes. This contradicts the specifications for the HP drive which has 256-Byte sectors.
- Each information block seemed to contain data (a few non-zeros) which is different for each block. This information seems to be created during the low-level formatting of the disk.
- Some data blocks contain information (significant number of non-zeros).
- If the data block contains information, it was presumably put there later (after the low level formatting). There are tiny distortions visible at the end of the information block. It is nearly impossible to perfectly synchronize with the original recording [13]. This creates tiny distortions which are clearly visible (in the cases evaluated). They are additional indicators for data written.

As some of these results contradicted the expected results, e.g. recording format and sector size, continued analysis was necessary.

Statistical analyses on the KryoFlux streams established the numbers for each length of signal (ground truth). There were visible clusters and only a few outliers (magnitudes less than the number in clusters). The distortions in sector writing are some of them. To reproduce data from the stream requires a correct clock signal, which is challenging as a binary “1” is a changing phase in the middle of a cycle (and then no additional clock signal is inserted); otherwise a clock is reproduced from mandatory phase changes of consecutive binary zeros.

```
# snippets of Python code for KryoFlux stream analysis
# distance of Data or Clock bits in M2FM multiples of 24
def genBitList() :
  for i in xrange( 300 ) :
  s = '1'
  i -= 24
  while i > 24 :
    s += '0'
    i -= 24
  if i >= 16 : s += '0'
  bitList.append(s)
# extract a byte 'Clock' and 'Data' of a byte
def pDataTakt(header , bits ) :
  # parse on BYTE DATA and CLOCk
  Clock = 0
  Data = 0
  for i in xrange(8) :
    Clock *= 2
    Data *= 2
  if bits[(i)*2] == '1' : Clock |= 0x1
  if bits[(i)*2+1] == '1' : Data |= 0x1
  Clock = revOrder(T)
  Data = revOrder(D)
  return revBitOrder(Data),revBitOrder(Clock)
# extract a complete sektor from bits
def pDataTaktData( header , bits ) :
  C = []
  T = 0
  D = 0
  c0 = ' 'c1 = ' '
  for i in xrange(0,len(bits),16 ) :
    C.append(bits[i:i+16])
    Sektor = []
    for i in xrange( len(bits)/2 ) :
      T *= 2
      D *= 2
    if bits[(i)*2] == '1' : T |= 0x1
    if bits[(i)*2+1] == '1' : D |= 0x1
    if (i%8) == 7 :
    c0 = B(c1)
    c1 = B(chr(revOrder(D)))
    Sektor.append(revOrder(D))
    ...
```

Taking this information into account and introducing a clock marker into the visualization (Figure 7), it became clear that some of the visible code violations do not match the FM possibility. Additionally, the statistical analysis on the phase-length FM could not completely explain the reproduced pattern. We now concentrated on the M2FM

encoding possibility and put out a call for help on Internet fora and to other universities.

Our next step was to create an actual bit stream for each data sector and read the sector information. The main challenge in this lay in the reassembling of the bitstream into bytes.

Reassembling the Bitstream

At a well-timed moment, we received responses to our call for help which provided us with more information for a decoder and confirmed the recording format as M2FM. With this knowledge, we started decoding the header information (short "sectors" after a synchronisation pattern) from the KryoFlux output.

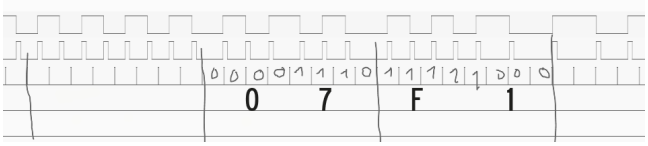

Figure 7: Extract of bitstream showing identification mark and track number

Sector information is not easily retrieved, as a track recording does not necessarily start with sector "0" and usually contains more than a single rotation of the platter. This information is contained in the short block after a sequence of 32 binary ones. We did a visual count in the representation from the logic analyzer software (see Figure 7). Directly after the pattern follows the identification mark where we found the 0x70 hex address mark (after bit swapping and reading from right to left) and then the track address information.

The track (or cylinder) number 31 (0x1F hex) was displayed in the next byte. Both bytes produced expected results which validated the approach. Having established the proper byte boundaries, the flux stream interpreter could now be completed. The simple structure of the disks simplified the task, as no complex hierarchical filesystem structure had to be taken into account to reassemble the files.

The proof for a properly working script was the reproduction of file names and occupied sectors which could be successfully compared to the sheets available in some of the disk sleeves.

```
ETTX..<.........  | TEXT...<.........   TEXT   60 - 61
OPEW.R=./.......  | POWER..=./......   POWER  61 - 108
OR.M.............  | ROM..............   ROM    925 - 1115
RBTASI..K.......  | BRATIS...K......   BRATIS 277 - 352
NIHFLD[.w.......  | INFHDL.[.w......   INFHDL 1115 - 1234
HCBKLE..\.......  | CHKBEL...\......   CHKBEL 465 - 557
...
```

During the different experiments, more facts about this floppy disk format were gathered. The disks contain an unusual number of sectors compared to the typical structure of IBM- and PC-style disks: 30 sectors instead of the expected 26. These sectors are interleaved and track numbers are not consecutive. The flux readings, created by the KryoFlux tools, contain in their track readings approximately five rotations of the platter. This offered the opportunity to compare different readings of the same sector to each other, and help eliminate errors due to inconsistent timing (jitter).

8 CONSOLIDATION

After three months of research, as well as trial and error experimentation, we received confirmation that the machine was an HP9845 with an external floppy drive, the HP9895A disk drive, attached. The disks have 77 tracks or cylinders, and contain 30 sectors with an interleaving factor of 7. The machine was marketed for scientific purposes and was popular in the university till the mid-1980s before the IBM PCs were introduced. It offered a wide range of peripheral connections and general programmed input/output (PIO).

Many of the floppy disks contained in the set appeared to have mostly data of clinical trials on them, judging from the file names and brief descriptions on the disk labels. At this point, we still did not know how these data were organized and if there was text contained within the files. The reason for looking for text files was

```
19514
19515  Sektorlength: 256
19516  57 20 72 65 65 74 65 20 77 72 72 61 65 74 20 6e    W reete wrraet n  |  Werte erwarten
19517  41 28 62 62 32 2e 2e 29 45 20 6e 69 20 65 6c 50    A(bb2..)E ni elP  | (Abb.2). Eine Pl
19518  74 61 61 65 62 75 6c 69 75 64 67 6e 73 20 74 65    taaebuliudgns te  | ateaubildung set
19519  74 7a 69 20 20 6d 6f 4e 6d 72 6c 61 65 62 65 72    tzi  moNmrlaeber  | zt im Normalbere
19520  63 69 20 68 65 64 65 72 65 23 e2 09 05 06 de 49    ci hedere#.....I  | ich dere#e....I.
19521  09 00 65 23 ec 09 2d 06 de 49 10 00 80 4d 68 58    ..e#..-..I...MhX  | ..#e...-I...M.Xh
19522  49 ff 20 20 20 20 20 20 20 20 56 20 32 4f 4b 2f    I.       V 2OK/   | .I        VO2/K
19523  20 47 69 65 2e 6e 46 20 65 75 20 72 65 50 73 72     Gie.nF eu rePsr  | G ein. Fuer Pers
19524  6e 6f 6e 65 6d 20 74 69 65 20 6e 72 65 69 72 64    nonem tie nreird  | onen mit erniedr
19525  67 69 65 74 20 72 61 6d 69 78 61 6d 65 6c 20 72    giet ramixamel r  | igter maximaler
19526  61 53 65 75 73 72 6f 74 66 66 63 2d 65 23 f6 09    aSeusrotffc-e#..  | Sauerstoff-c#e..
19527  05 06 de 49 09 00 65 23 00 0a 2d 06 de 49 10 00    ...I..e#..-..I..  | ..I...#e...-I...
19528  80 4d 68 58 49 ff 20 20 20 20 20 20 20 61 20       .MhXI.        a   | M.Xh.I        a
19529  66 75 61 6e 6d 68 20 65 65 62 74 73 68 65 20 74    fuanmh eebtshe t  | ufnahme besteht
19530  69 65 65 6e 73 20 72 74 6e 65 20 67 69 6c 65 6e    ieens rtne gilen  | eine streng line
19531  72 61 20 65 65 42 69 7a 68 65 6e 75 20 67 77 7a    ra eeBizhenu gwz  | are Beziehung zw
19532  81 31                                              .1                | 1.
19533
19534  CRCDATA:  0x0
```

Figure 8: Decoded text from an 8-inch disk

that text was encoded in ASCII with each character taking one byte of storage, while the numeric data was encoded with different byte sizes dependent on whether they were full-precision numbers, short-precision numbers or integers. One of the disks was labelled "Publications", so this was selected to test our flux stream decoder. An extract of the result of decoding one track is in Figure 8, which shows the hexadecimal value of each byte, the ASCII character, and the byte-swapped version.

9 IMPLICATIONS

Preservation projects that we and others have undertaken have shown us that legacy hardware is still a significant part of the digital universe [7, 8, 13, 23]. These projects may be in the form of complete systems or (removable) storage media. The last 30 years produced a plethora of magnetic, optical and more recently electronic removable and fixed media including audio and video cassettes, cartridges, floppy disks, ZIP disks, Syquest drives, hard disks, and so forth [11, 19], in a variety of formats. Many of them were deployed as archival media, especially the cheaper ones. While floppy disks and optical media such as CDROM, DVD (or Blu-ray Disks) might be more common, other items including ZIP, Syquest, Jaz, LS-120, SparQ, Orb or MicroDrive might surface in the legacies delivered to special collection libraries or digital arts institutions. More recently, electronic media like CompactFlash cards, memory sticks or multimedia cards will complement the already wide selection.

The drives to handle certain removable media have begun to disappear from the typical second-hand markets, e.g. it is much more difficult to obtain a (confirmed working) 5.25-inch drive than

it was just five years ago. In short *Successful [media] migration diminishes as the age of the medium, or hardware necessary to read it, increases. While most studies focuses [sic] on the longevity of the medium, no doubt fuelled by manufacturing marketing, the true risk lies in the scarcity of hardware necessary to read these formats* [10].

Compared to the ease of low-level analysis of 8-inch media, the challenges will increase for most of the recent much more compact and integrated storage peripherals. Some devices became much more proprietary and were manufactured by just one company. These developments increased the number of different standards, but narrowed the amount of different media available for a particular device. Increasing capacities require more advanced technology to squeeze data onto increasingly smaller media.

For many smaller institutions it does not necessarily make sense to maintain a full collection of possible readers and peripherals. For a range of once-popular formats, business models are well established like media transfers from various magnetic tapes. Additionally, there is a selection of transfer peripherals available as listed in the addendum of [13]. Nevertheless, some implementations, especially for USB, might be incomplete or erroneous.[1]

One strategy, especially for digital art collectors and memory institutions, is to buy spare older equipment for backup or replacement parts, filling their storerooms and taking precious space without the guarantee of usefulness when required [2]. But, as our studies confirmed, there are certain gaps in the service domain challenging memory institutions, such as Freiburg, Archives New Zealand and others seeking to read old 8-inch disks.

Bridging the Technical Gap

A digital object is created in a physical and system environment [22]: the former with specific hardware components, the latter with a specific operating system version and utilities. Specialised software, such as digital art and video games, depend on particular hardware for interaction or rendering. Peripherals and controllers need hardware interfaces which become rare as computer technology evolves. While the development of system emulators is seen as a solution to some of these issues, original hardware is needed to prove that an emulator is working properly.

Preservation of born-digital objects necessitates links between past, present and future hardware. A great deal of this can be handled via peripherals. As demonstrated in this 8-inch floppy disk case, challenges are faced at different stages and technical levels. While some of these challenges are unique to each specific type of medium, others follow a pattern, and these will surely repeat for other storage media, such as the following:

- A peripheral to read a particular storage medium must be powered in a specific way. Previously, the voltage and power requirements were much higher than today, 24V and alternating current (AC) connections are now uncommon. In addition, the old types of connectors are no longer manufactured and their documentation of the pin assignment does not typically come with the device.
- To transfer the data to new media, a peripheral or complete computer system is connected to a current system. In many cases, the interfaces do not exist or have changed significantly. An 8-inch drive had a 50-pin connector; for the later 5.25-inch and 3.5-inch drives, these were replaced with 34-pin connectors.
- Parallel connections were superseded by high-speed serial connections as seen in the transition of IDE to SATA.
- The Universal Serial Bus (USB) standardised the connection of commonly used computer peripherals to personal computers, both to supply electric power and to communicate. Thus, it is possible to connect a RS232 serial cable or IDE and SATA cables to USB via an interface adapter. However, the voltage supplied by the USB may not be high enough for older serial connections, and some IDE adaptors do not comply with all aspects of the specification standard making it difficult to properly image e.g. an early laptop IDE hard drive.
- Lack of technical knowledge about connectors, unusual sockets, and plugs cause difficulties: incorrect wiring can destroy well functioning, precious hardware.
- Small embedded components or a System-on-a-Chip (SoC) can replace complete controllers, nonetheless at some point a particular cable or power supply will still be required, as well as the knowledge and expertise on how to connect it.
- Finally, some "glue component" like a firmware or driver might be required to make the peripheral accessible from the operating system.

Hardware and Documentation Repositories

The need for preserving hardware, software and documentation has been raised for decades. In 1992, the Commission on Preservation and Access [15] recommended that *Computer museum operators can be urged to maintain software as well as hardware, and to be able to operate old programs for purposes of translation.* The same call was made by McDonough et al. in 2010 [18] and Rieger et al. in 2015 [21]. As Buskirk [2] put it *... equally urgent is burgeoning obsolescence that leaves orphaned media with no equipment on which to be played or software that is incompatible with current platforms.* This aim has not yet been realised.

Dissemination of preservation techniques and workflows is well established for tasks to be undertaken *after* a usable digital image is acquired, and for popular formats, e.g. 8mm film or VHS to DVD, but little is available for older or rarer artefacts. Much of what is available is in the hands of hobbyists, enthusiasts and digital forensic scientists [10]. Although many hobbyists provide information online, these sites are often not updated and if interest in the topic wanes, the effort needed to maintain the site may outweigh the use for the site owner(s). During this project, the website on which we found the information for the Nashua disk disappeared. The National and State Libraries of Australasia [24] report that inhibitors to data transfer include

- no in-house equipment and equipment is not easy to source
- lacking the appropriate drives
- no internal expertise or staff that have suitable skill set
- no robust workflows for disk imaging

This paper has concentrated on 8-inch disks in M2FM format; data retrieval methods for BTOS/CTOS formatted disks at Archives New

[1] Challenges and solutions of imaging 20-plus-year-old IDE disks, http://openpreservation.org/blog/2013/03/20/challenges-dumpingimaging-old-ide-disks

Zealand and the University of Freiburg are documented in [23]. Substantial personnel effort and time is invested in setting up the equipment for these projects, as well as financial costs. Often, the equipment is used for a single project, meaning that the return of investment is low, and the acquired knowledge is lost over time.

There are many computer museums around the world, however their primary mission is to preserve and display computer systems, not the preservation of born-digital objects. There is a growing need for there to be centres for the transfer of born-digital artefacts to current media which are hubs of expertise and resources, who have a collection of obsolete computers, peripherals, software and manuals, such as the National Library of Australia, for their digital collection, and the Computer Archaeology Laboratory at Flinders University, who undertake research into recovering data from obsolete media.

Technical documentation is a vital resource which is also rapidly disappearing, hardcopy manuals and handbooks are seldomly digitised for online access, especially for devices that are currently perceived by the general public to be out of use.

10 CONCLUSION

Preservation of data from obsolete digital media is not a straightforward exercise. While many floppy disk formats have been studied and have solutions available for them, there are still many, which were widely used, that have not yet been deciphered. This project took a lot of patience, a lot of following hints and hunches to track down relevant information, and experimental analysis. We were also aided enormously by the generosity of strangers.

Confirming the disk format and model of the original system took quite a while, until we finally made contact with someone who was involved with the scrapping of the system. Various different levels of forensic science came into play: physical, logical, and higher level information were all needed to identify the system, the drive, the disk format so that we could finally extract some text.

Extracting other data content is much more complicated. We now know that mostly scientific analysis work was carried out on the system, with a large amount of the data being used in databases, spreadsheets and custom-written analysis software. It is difficult to interpret the non-textual data without the context or the software, and we cannot tell if we have the right numeric values, whereas it is easier for ASCII text. We pretty much ended at the point of the 5.25-inch floppy disk recovery study being in need of the original software of the system the data was created with [23]. Additionally, it could be helpful to directly link KryoFlux images to special stream readers built into emulators that interpret the streams directly as disk drive sources, such as in the WinUAE Amiga emulator.

As seen in Figure 8, there are undecipherable characters within the text, these could be word processor formatting codes for specific styles or characteristics of the document, or something else entirely.

Our experiences have reinforced the importance of and need for repositories of knowledge and resources for the initial steps of digital preservation – the creation of usable digital images – which should parallel the sharing of knowledge that has evolved around the handling and future-proofing of artefacts held in galleries, libraries, archives and museums. Especially smaller institutions, lacking the resources for such special-purpose departments in-house, would benefit significantly from these repositories. Such activities must be complemented with modified archival workflows to prevent items falling into limbo between becoming out of use within the institution and being handed over for archiving. Research data management will help to mitigate such challenges in the science domain.

REFERENCES

[1] BitScope. 2016. *BitScope Micro.* http://www.bitscope.com.

[2] Martha Buskirk. 2014. *Bit Rot: The Limits of Conservation.* Hyperallergic. http://hyperallergic.com/131304/bit-rot-the-limits-of-conservation.

[3] David Collins. 2016. *HP Computer Museum.* http://www.hpmuseum.net/.

[4] Jason Curtis. 2017. *Museum Of Obsolete Media.* http://www.obsoletemedia.org/8-inch-floppy-disk/.

[5] D Bit. 2016. *FDADAP floppy disk adapter.* http://www.dbit.com/fdadap.html.

[6] Bart de Vries. 2016. *Digital Obsolescence: Reproducing Floppy Data Cables.* http://openpreservation.org/blog/2016/09/02/digital-obsolescence-reproducing-data-cables.

[7] Denise de Vries and Craig Harrington. 2016. Recovery of heritage software stored on magnetic tape for Commodore microcomputers. *International Journal of Digital Curation* 11, 22 (2016), 10.

[8] Denise de Vries and Melanie Swalwell. 2016. Creating Disk Images of Born Digital Content: A Case Study Comparing Success Rates of Institutional Versus Private Collections. *New Review of Information Networking* 21, 2 (2016), 129–140.

[9] Jürgen Enge and Heinz Werner Kramski. 2016. Exploring Friedrich Kittlerfis Digital Legacy on Different Levels: Tools to Equip the Future Archivist. *iPRES 2016* (2016).

[10] Miriely Guerrero. 2012. Division Records-Digital Curation-Papers-Removable Media and the Use of Digital Forensics. (2012). https://deepblue.lib.umich.edu/bitstream/handle/2027.42/96441/Guerrero_IMLS_RemovableMediaReport_20120702.pdf Bentley Historical Library.

[11] David Christopher Harrill and Richard P Mislan. 2007. A small scale digital device forensics ontology. *Small Scale Digital Device Forensics Journal* 1, 1 (2007), 242.

[12] John F Hoeppner and Larry H Wall. 1980. Encoding/Decoding Techniques Double Floppy Disk Capacity. *Computer Design* 19, 2 (1980), 127–135.

[13] Matthew G Kirschenbaum, Richard Ovenden, Gabriela Redwine, and Rachel Donahue. 2010. *Digital forensics and born-digital content in cultural heritage collections.* Council on Library and Information Resources.

[14] Ansgar Kueckes. 2016. *The HP 9845 Project.* http://www.hp9845.net.

[15] Michael Lesk. 1992. *Preservation of New Technology. A Report of the Technology Assessment Advisory Committee to the Commission on Preservation and Access.* Commission on Preservation and Access, Washington DC, USA.

[16] Jean Louis-Guerin. 2013. *KryoFlux Stream File Documentation.* www.kryoflux.com/download/kryoflux_stream_protocol_rev1.1.pdf.

[17] Tim Mann. 2010. *cw2dmk.* https://github.com/ezrec/cw2dmk.

[18] Jerome McDonough, Robert Olendorf, Matthew Kirschenbaum, Kari Kraus, Doug Reside, Rachel Donahue, Andrew Phelps, Christopher Egert, Henry Lowood, Susan Rojo, and others. 2010. Preserving virtual worlds final report. *University of Illinois at Urbana-Champaign* (2010).

[19] National Library of Australia. 2008. *Mediapedia: Physical Format Carrier Resource.* http://mediapedia.nla.gov.au/.

[20] DPC Online. 2017. *Digital Preservation Handbook Legacy Media.* http://www.dpconline.org/handbook/organisational-activities/legacy-media.

[21] Oya Y Rieger, Tim Murray, Madeleine Casad, Desiree Alexander, Dianne Dietrich, Jason Kovari, Liz Muller, Michelle Paolillo, and Danielle K Mericle. 2015. Preserving and emulating digital art objects. (2015).

[22] Kenneth Thibodeau. 2002. Overview of technological approaches to digital preservation and challenges in coming years. *The state of digital preservation: an international perspective* (2002), 4–31.

[23] Dirk von Suchodoletz, Richard Schneider, Euan Cochrane, and David Schmidt. 2012. Practical Floppy Disk Recovery Study. *Preservation of Digital Objects* (2012), 184.

[24] Scott Wajon, Somaya Langley, and Damien Cassidy. 2016. *Collaboration on Digital Infrastructure Project 3: Obsolete Physical Carriers in NSLA Collections.* National and State Libraries of Australasia. http://www.nsla.org.au/publication/obsolete-physical-carriers-nsla-collections-stage-1.

[25] Webresource. 2005. *Fdutils.* https://fdutils.linux.lu/disk-id.html.

[26] Wikipedia. 2016. *IBM 8-inch formats.* https://en.wikipedia.org/wiki/List_of_floppy_disk_formats.

ACKNOWLEDGMENTS

Many thanks to
Mick Crouch (Archives New Zealand) and Euan Cochrane (Yale University Library) for advice and moral support,
Eric Smith for information about HP devices,
Christian Corti (Computer Center University of Stuttgart) for assistance with analyzing the M2FM encoding,
Konrad Meier (Computer Center University of Freiburg) for providing the oscilloscope and helping with the configuration of the experiment and interpretation of the readout, the Freiburg Emulation-as-a-Service team to provide support and hardware, and finally Jeannette Vollmer to proof read the paper.