6-2016

Working with Legacy Media: A Lone Arranger's First Steps

Elizabeth Charlton

The University of Paris I - Pantheon-Sorbonne

Working with Legacy Media: A Lone Arranger's First Steps

Description
[Excerpt] In 2013, a naked hard drive from Fiji arriving in my small religious archives (an equivalent full-time staff of 2.5 – one archivist and two archives' assistants) started me off on the path of digital preservation and, in particular, the digital forensics practices that are beneficial for archivists. With such a small staff, outsourced IT services, and no digital preservation policy in sight, it was time to start exploring how institutions of my size could manage legacy media and start planning for the born-digital archives that will continue to arrive. Since I hold a part-time position, I was able to undertake this exploration in my own time through the support provided by a scholarship from the Ian McLean Wards Memorial Trust in 2015.

Keywords
digital collections, legacy media, preservation

Working with legacy media: A lone arranger's first steps

Elizabeth Charlton

The University of Paris I – Pantheon-Sorbonne

Background

In 2013, a naked hard drive from Fiji arriving in my small religious archives (an equivalent full-time staff of 2.5 – one archivist and two archives' assistants) started me off on the path of digital preservation and, in particular, the digital forensics practices that are beneficial for archivists. With such a small staff, outsourced IT services, and no digital preservation policy in sight, it was time to start exploring how institutions of my size could manage legacy media and start planning for the born-digital archives that will continue to arrive. Since I hold a part-time position, I was able to undertake this exploration in my own time through the support provided by a scholarship from the Ian McLean Wards Memorial Trust in 2015.

Introduction

Due to this context, and the limited level of my own knowledge, I started an incremental process, taking small steps, to get on the board for digital preservation. One of them was attending DigCCurr Professional Institute where during the first session, Nancy McGovern posed the question, "What is good enough practice in your situation?".[i] Keeping this question in mind framed, contextualised, and directed my subsequent enquiry.

For the scholarship application, I determined the following targets: learning how to get material off legacy media, monitoring whatever is retained for fixity, instituting a technology watch and putting policy in place. The quantity of material in my institution

currently requiring management is relatively small – a small box of mostly unidentified floppies, CDs, a naked hard-drive, and two laptops. Furthermore, the standalone archives' computer had died before my employment, so I removed the hard drive to ensure that any necessary business records for the running of the archives were retained. It is congregation policy that the Archives receive the personal papers and effects of deceased members for appraisal and disposal. Looking toward the future, with an ever-aging congregation I expect to receive on average one laptop or naked hard drive a year. It is also possible that the administration's shared drive may end up coming my way.

Collaboration is an important factor in digital curation.[ii] An additional output from my scholarship research was to create a workflow and self-service instructions for disk imaging with the equipment selected for the project. Guidance to undertake some very basic processing tasks in order to gain enough information to appraise the contents and prepare the material selected for retention was outlined. The motivation was to share the equipment with archivists of similar-scale institutions with whom I have regular contact.[iii]

My institution is a member of a long-standing informal group of lone arrangers, a group which meets regularly for site visits and sharing. Members include archivists and volunteers in schools, business, community and religious groups. I sent out a survey with the monthly newsletter to ascertain the types of physical media found in their collections. Their media encompass an external hard drive, CDs, flash drives and unidentified 3.5" floppy disks. Some colleagues are running oral history programs. They started recording on cassette tapes and currently use SD cards, so they will need to know how to appropriately manage this born-digital material. Some colleagues identified that they do not have strong computer literacy skills. It was imperative to provide them with guidance so that they too could start managing their legacy media independently and improve their management of born-digital content.[iv] In order to create this material, I needed to go through the process of disk imaging and the

subsequent pre-ingest processing myself so that I would be in a position to provide clear, explicit instructions.

Some American digital archivists maintain it may be better for small institutions to outsource these processes.[v] Erway, Goldman & McKinley produced a paper detailing the requirements for outsourcing legacy media processing.[vi] At this stage in the New Zealand context, it is unlikely that the audience that I have targeted would have the financial support to contract someone to complete this work – my own institution certainly does not – one reason to share equipment and knowledge with them. Furthermore, given the size of the cultural heritage sector, it would be unlikely that this type of service would be financially viable unless it was a unit of a bigger organization; staff with the requisite knowledge would need to be found and most probably would not initially have a fulltime workload. To date, there are only two cultural heritage institutions that have staff dedicated to working with legacy digital material, the National Library of New Zealand and Archives New Zealand; they do not offer user-pays services. There are a number of companies who offer digital forensic services, though they focus on current media, not the legacy media found in archives; only one specifically mentions imaging floppy disks.[vii] Given that most of the floppies in my colleagues' and my institutions have no recorded provenance, it would not be viewed as an appropriate use of our very limited budgets to outsource the processing since the key aim of this stage is to appraise the material. Apart from this issue, carrying out at least some form of legacy media processing would provide a fantastic learning opportunity to archivists in small institutions, to upskill and to stimulate awareness and conservation about digital preservation and curation.

Initial Scope and First Steps

One of the principles in digital preservation is not to create any irreversible changes to the data and if any were to occur to document them.[viii] If I were to make any changes while learning, I wanted data that was not of any long-term value.

I selected to start by practicing imaging 3.5" floppy disks as this was the most numerous media indicated by my colleagues as well as in my own institution. As I had 31 3.5" floppies retained from my teaching days in Australia, the disks fulfilled my criterion of not being of long-term value. From what I could remember, these floppies were in use around 1997-2002. I expected that they contained end-of-unit tests and grammar notes for classes I taught at the time. I also expected that these were just transfer media and that I would find copies of most, if not all the data on an external hard drive copied from my Australian desktop computer.

Besides, the size of a floppy means it images quickly and the number I had to hand offered the repetition conducive to embedding new learning. As I describe below, disk images were invaluable when the floppies themselves would not mount.

Establishing a workflow for appraising material found on legacy media was one of the goals of imaging my floppies. The collecting scope of a congregation archive is relatively narrow, with the added dimension of its being subject to Canon Law. This brings with it the requirement to retain whatever documents the temporal and spiritual affairs of the congregation. If the material does not record the congregation's story in New Zealand or of the overseas ministry of New Zealand members, then it is out of scope, so it is important that material is appraised.

The workflows presented in Marty Gengenbach's thesis appear to indicate a known provenance for the material being processed.[ix] In the case of the floppies at work, they have been separated from their original collections without this being recorded, hence my emphasis on processing for appraisal. Since I did not know how they have been stored, I worked from the assumption that it would be possible to access them only once – another motivation to make an image to capture whatever it contains. Following a standardised process assists in building confidence and provides a structure to follow whatever the type of media to be processed. While my backlog of material may be small, the longer it is left the greater the challenge will be to source suitable equipment and have readable media in order to appraise it.

Wilsey, Skirvin, Chan, & Edwards use the phrase "imaging philosophy" as a section heading for the following:

Before imaging, we needed to think carefully about our approach. Should we attempt to capture every bit of every removable storage media? Would we be satisfied with imaging most of the media? In the end we decided to attempt to capture media that was readable by Mac or PC computers with the hardware (such as external floppy disk and Zip drives) that we already had in place. We chose to attempt reading each disk, but did not attempt to diagnose why a particular disk wasn't readable. Under a tight deadline, we chose to follow this philosophy so that we could image as many readable disks as possible without spending too much time on troubleshooting.[x]

I found this concept to be a pertinent consideration and one that is not unique to digital material. For example, in our collection there are notebooks written over a hundred years ago in shorthand by a Frenchman. At first glance it is unclear whether it is an English-based, a French-based or a personalised shorthand, then the language will also need identifying. How much time should be spent on identifying the shorthand and decoding it?

So when does spending too much time on troubleshooting start? What springs to mind immediately is the cost consideration if the disk does not image on the first go. How many different drives are required to be able to say that we attempted to read the disk? Literature and face-to-face discussions point out that a drive that works for one disk may not for another; that it can be a case of trial and error.[xi] How much time do you spend before giving up? If you are unsuccessful what do you do with that media – deaccession, keep?

What happens when you have successfully imaged the media, but in order to evaluate the value of the contents and to prepare it for ingest into a digital repository, even more detective work is required? In a general policy, how specific should one be

about one's imaging philosophy? Or should the imaging philosophy change according to the provenance of the media being worked on? In fact, should this concept be taken even further to encompass the processing that needs to occur post-imaging, so that question becomes, "What is your processing philosophy"? As I worked through my test corpus I hoped this exercise would assist me in resolving some if not all of these questions.

Preparing for Imaging

I followed the guidelines in the OCLC report, You've got to walk before you can run: First steps for managing born-digital content received on physical media, to set up a spreadsheet in order to create an inventory of the floppies in my possession and to record the results of the imaging process.[xii] As seen in Figures 1 and 2 below, very few of the floppies had useful labeling.

An early topic of discussion for archivists was the choice of image format, as an image itself is a file. I selected to image in the proprietary Encase format, E01, as Simson Garfinkel has stopped maintaining the AFF format since Encase was reverse-engineered.[xiii] Other institutions have selected to retain their images in the raw format.[xiv]

In order to image, that is, to create an exact replica of the contents of the source medium – data and structure contained in a single file, I used an off-network 17.3" Toshiba laptop with an i7 processor, 16GB RAM, 1TB hard drive, running Windows 8.1, loaded with the tools introduced at the Society of American Archivists Digital Forensics: Advanced course, which are open-source and freely accessible from the Internet. The two programs used in particular for this exercise were FTK Imager and BitCurator 1.5.7. Since BitCurator is built in Ubuntu, it was accessed through VMWare. Laptops no longer come equipped with internal floppy drives so two USB drives were used, branded TEAC and Mitsumi.

It is usual practice to use a write blocker to protect data from any changes. A write blocker is a mechanism that does not allow anything to be written to the media. I had read that USB-driven disk drives do not usually work when attached to a USB write blocker and it was preferable to use a Digital Intelligence floppy drive, which had an inbuilt write-blocking mode.[xv] However, by the time I found out about them, I could no longer source one. Unlike other media, floppies came with their own built-in write blocker, the read-only tab so I ensured that the tabs were correctly positioned before inserting any floppies into the drive. BitCurator includes a software write blocker as a standard feature, providing an extra level of security.

Figure 1 – My collection of floppy disks numbered from left to right (Disks 1-24)

Figure 2 – Disks 25-31

Disk images of the floppies were created three times – twice using Guymager in BitCurator with both drives and once with FTK Imager with the TEAC drive. Waugh's study on floppies had been conducted with FTK Imager, prior to the formal release of BitCurator, and I wanted to see if I had the same issues with the former software.[xvi]

Once all the images had been made, they were run through ClamTK, the virus checker included in BitCurator, before passing the BitCurator Reporting Tool over them. The contents of the images were viewed through BitCurator Disk Image Access and GHex in BitCurator or FTK Imager in Windows to make appraisal decisions.

While I would not do this with media being evaluated for long-term preservation as it may modify data, for curiosity sake, I also tried mounting the physical disks to see if I could access them. Mounting is the process that allows a computer's operating system to access external storage media.

Imaging Results

Waugh had a number of Macintosh/Apple format disks that had not been able to be read by the drives and software she had access to.[xvii] Because of this, I expected that the three disks (Disks 1, 25, 26) that had Macintosh/Apple format labels would not read on my set-up since I was imaging using a Windows-operated machine and that I would need recourse to a Kryoflux, a floppy drive controller and additional floppy drives; however, to my surprise, I was able to image all disks with both drives. Disk 1 labeled as an Apple diskette had in fact been reformatted to Windows.

Upon mounting the images, eight were found to be empty – there were no files under root when the file directory was expanded, the unallocated space measured 1.42MB and the empty image files were all sized 12KB; this was recorded in the spreadsheet and the images were discarded.

Exploring processing tools

With the successful imaging of all the disks, I was able to move on to the analysis of the images, using tools provided in FTK Imager and in the BitCurator environment. Even though I thought I had reasonable knowledge of the disk contents, I wanted to go through the processes I would undertake were I confronted with unknown disks, as I will

be at work. Colleagues will be in similar situations so I wanted to be able to advise them as well.

Extracting files

As an image is an exact replica of the source medium, there may be times when some of the files need to be accessed. In this case, it is possible to extract files from the disk image.

In FTK Imager, I was able to view the images' file trees, which provided me with enough information about their contents. I was able to export files directly from the image. In particular, this direct export was handy with the three disks of the Mac HFS file system (Disks 17, 25, 26). HFS was in use from 1985 and was finally superseded in 1998.

Within the BitCurator environment, HFS Explorer has been included as the Sleuth Kit upon which the reporting tools rely, does not process Mac file systems prior to HFS+ (i.e. MFS and HFS). However, HFS Explorer requires an image in raw format so an extra step of exporting the Encase format to raw is required.[xviii] Once accomplished, I was able to view and extract files. Windows and later Mac images are viewed using BitCurator Disk Image Access.

Individual Mac files extracted by both programs keep the original modified date, whereas Windows files extracted by BitCurator indicate the date of extraction. It is possible to extract the whole file structure in FTK Imager, while the folders created show the date of extraction, the files keep the original modified date. Retaining original timestamps assists, of course, with documenting a file's authenticity and integrity. Jarrett Drake mounts the image and writes a script that allows for transfer without changing the metadata.[xix] When this is not possible, a work-around could be to use the snipping tool or take a screenshot of the timestamp shown within the image and save this.

Format identification

Five disk images were found to contain files requiring further investigation in order to verify their format. Format identification is one aspect of characterization, the process of collating information about the structure and content of a file. This needs to occur before ingesting the data into storage. Since preservation is for use, it is important to ensure that a minimum of information required for preservation actions and to maintain meaningful access over time is present. The greater the time from the date of creation, the more complicated becomes the collection of information. Characterization can also include format validation and metadata extraction.[xx] Many of the more advanced processes of characterization are automated and integrated into digital preservation software. The use of such software was out of the scope of the first steps I set out to achieve.

One disk image (Disk 20) had an unidentified file structure, though it was clear from its properties that it was not blank. This disk was imaged with 201 bad sectors, including a corrupted boot sector, which led to the unidentified file structure. Viewing the image in a hex editor, I was able to find the signature for a Word document.

Three disk images (Disks 25, 26, 31) contained files with no file extensions. I used these files to practice format identification. Two were the HFS images (Disks 25, 26). Judging by the file names, I expected them to be all Word documents. Viewing the file signature in a hex editor confirmed that this was the case for most of them; the remainder were in rich text format. For one file extracted from Disk 26, I also added the .doc extension by using the file rename feature before attempting to open it, ran the file through Droid, the National Archives file format identification tool, as well as the hex editor check.[xxi] All three methods were successful in their format identification.

Disk 31 contained two files (logo, logo2) with no extension. These files were found within resource.frk, indicating that the items had been copied from a Mac. Again using the hex editor, JFIF was seen in the text viewer. By checking the Pronom file

format registry, the start and end strings for a JFIF were identified.[xxii] They were found within the file, and the data between them copied into a new hex editor window. This was then saved. Adding the .jpg extension brought up the actual logo, that of the French Teachers' Association of Victoria.

The unexpected HFS disk (Disk 17) contained three files with the extension .tlm. This extension is identified as Timeline Maker.[xxiii] I had never heard of that software. The date and name of the files (boys' first names), and viewing as text in a hex viewer, confirmed that the contents were to do with marks from a unit test. No headers were visible at the start of each file (Figure 3). Scrolling through the whole image did not bring any header to light. I extracted one of the files and added .doc, .rtf and .xls extensions; only the first word rendered in LibreOffice word-processing (Figure 4) and spreadsheet programs. The marks had been processed statistically using SSPS in Windows in 1999 to analyze the construction of multiple choice questions in a post-graduate paper on assessment. In fact, Disk 27 contained the data generated in SPSS in an Excel spreadsheet.

```
000 6E 69 63 6B 0D 31 32 0D-31 31 36 37 0D 5B 23 55  nick·12·1167·[#U
010 6E 69 74 31 3A 20 5B 31-3A 20 5B 31 2C 20 31 2C  nit1: [1: [1, 1,
020 20 30 2C 20 30 2C 20 31-34 33 5D 2C 20 32 3A 20   0, 0, 143], 2:
030 5B 31 2C 20 31 2C 20 30-2C 20 30 2C 20 31 30 36  [1, 1, 0, 0, 106
040 5D 2C 20 33 3A 20 5B 31-2C 20 31 2C 20 31 37 35  ], 3: [1, 1, 175
050 2C 20 31 37 35 2C 20 32-33 33 5D 2C 20 34 3A 20  , 175, 233], 4:
060 5B 31 2C 20 31 2C 20 31-37 35 2C 20 31 37 35 2C  [1, 1, 175, 175,
070 20 32 30 32 5D 2C 20 35-3A 20 5B 31 2C 20 31 2C   202], 5: [1, 1,
080 20 32 37 35 2C 20 32 37-35 2C 20 31 35 39 5D 2C   275, 275, 159],
090 20 36 3A 20 5B 31 2C 20-31 2C 20 31 30 2C 20 31   6: [1, 1, 10, 1
0a0 30 2C 20 31 34 33 5D 2C-20 37 3A 20 5B 31 2C 20  0, 143], 7: [1,
0b0 31 2C 20 35 31 2C 20 35-31 2C 20 31 36 33 5D 2C  1, 51, 51, 163],
0c0 20 38 3A 20 5B 31 2C 20-31 2C 20 39 36 2C 20 39   8: [1, 1, 96, 9
0d0 36 2C 20 32 30 31 5D 2C-20 39 3A 20 5B 31 2C 20  6, 201], 9: [1,
0e0 31 2C 20 31 36 30 2C 20-31 36 30 2C 20 31 32 36  1, 160, 160, 126
0f0 5D 2C 20 31 30 3A 20 5B-31 2C 20 31 2C 20 36 30  ], 10: [1, 1, 60
100 2C 20 36 30 2C 20 31 36-32 5D 2C 20 31 31 3A 20  , 60, 162], 11:
110 5B 31 2C 20 31 2C 20 31-30 39 2C 20 31 30 39 2C  [1, 1, 109, 109,
120 20 31 37 35 5D 2C 20 31-32 3A 20 5B 31 2C 20 31   175], 12: [1, 1
130 2C 20 35 36 2C 20 35 36-2C 20 33 37 33 5D 5D 5D  , 56, 56, 373]]]
```

Figure 3 – hex editor view of .tlm content
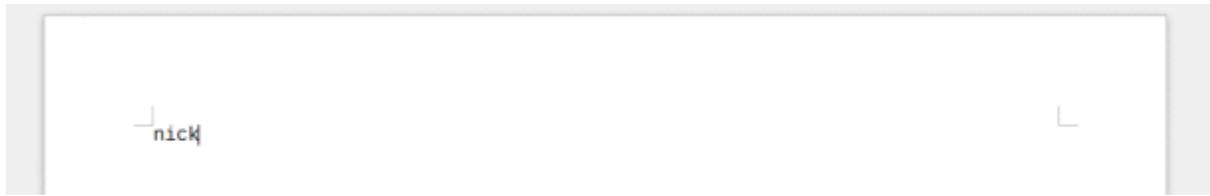
Figure 4 – .tlm file rendered as .rtf

At DigCCurr's winter session, it was suggested that the numbers between the square brackets could indicate coordinates in a graph. Since I still had all the working files for the university paper on my external hard drive, I viewed the Excel files in the hex editor in FTK Imager to determine whether they presented in a similar fashion. They did not. Though doing this did inform me that at some stage I had used a Mac with this material as a resource.frk folder was present.

This exercise in file format identification was very useful, though the most time-consuming of all the tasks trialled. For initial appraisal, the hex editor was most beneficial. For files requiring further work, such as the logos, which had uninformative file names, being able to render them meant that I could make an informed appraisal decision.

Repairing disk images

A bad sector is one that cannot be accessed or written on, so I was also surprised that the number of bad sectors sometimes changed according to the floppy drive I used. This can be caused by the different types of design of floppy drive heads.[xxiv] It was also suggested that I try running a disk repair tool over the image such as fsck in Linux, a command line tool.[xxv]

I acknowledge that many of the processes involved with pre-ingest require command line knowledge. I have personally found it challenging to decipher the instruction guides for some of the tools; it appears that they have been written with the

assumption that the people using them have a reasonable level of proficiency. While I understand the concept behind command line, I struggle to enter the commands correctly. With my target audience in mind, I looked for tools with a graphic user interface. Once I have finished this project I will put command line proficiency on to my goal list as I continue to incrementally develop my skills.

I found Photorec, an open-source application with a graphical user interface, which will repair files in a raw disk image.[xxvi] I extracted a raw image using FTK Imager from Disks 2, 5, 17, 20, 22, 25 and 26, which I then used. These disks were selected because they had either imaged with bad sectors or the physical disks themselves would not mount (see below). Some of the files, which I had initially identified through the hex editor, were recovered. The majority of the files could be opened and on the whole rendered well, excepting Disks 20 and 25. For Disk 20 the image showed an unrecognised file system as well as unallocated space. Photorec recovered four files of report comments, though there were some sections of text unable to be recovered. Most of the files on Disk 25—a Mac formatted disk—were recovered and rendered without issue. The exceptions were a couple of files, written in French and German, languages employing diacritics.

As mentioned earlier, I attempted to mount the physical disks. More questions arose when finding that eleven of them would not mount – three were the Mac formatted disks (Disks 17, 25, 26), one had the unrecognised file system (Disk 20), but there still remained seven Windows disks. Since "most disk imaging tools ignore file system data (they capture it, but don't require it to be intact)", this is a possible reason why I was able to image disks but not mount them.[xxvii] This experiment further strengthens the practice to image floppies straight away. Had I just relied on mounting the physical disks, material that I was able to access through the images would have remained inaccessible.

Discussion

These floppies were a "known" collection in that they belonged to me and I had created them. As the photographs show (Figures 1 and 2), most of the disks did not have labels and those that did were not helpful, e.g. "Please return to E. Charlton". That phrase, at least, bore out my contention that I had used them for transfer and not for storage.

I checked the hard drive containing files from the same time period. I found the contents of five disks (2, 3, 12, 17, 25) were not replicated on the drive. Using archival appraisal principles, I then considered whether any of the files found only on the disk images were worthy of long-term retention and preservation. Disk 17 has already been discussed for its unusual file format. Some of the contents of disks 2 and 3 were at one stage on my desktop computer. However, as my involvement with the organizations had ended, I had deleted those files before ultimately transferring the desktop's content to the external hard drive. My opinion of the contents was not changed upon review. Disk 25 was another Mac formatted disk; the content once again was of a transitory nature.

The only items which caught my attention from an appraisal perspective were the photos (Disk 12) of my Year 8 homeroom class from 2000. They were informal shots of the group. The size of the images (117kB, 120kb) meant that I decided that it would not be worth the resources to retain them into the long-term. Furthermore, I do have the official class photo, the school yearbook and the school has an archives, which contains photos of much better quality than these.

A comment that has come up regularly regarding disk images is that they capture EVERYTHING in the bit-stream.[xxviii] Files that have been deleted become accessible through imaging. Is this what the creator wants researchers to have access to? Seeing this occur with my own information brought this message home very strongly. For example, I was rather startled to see a list of email addresses appear in Bulk Extractor. Upon looking at the directory tree, I found Eudora mailboxes (.mbox and .toc) and

realised that I had used that floppy to copy the mailboxes from one computer to another. After transfer, I had deleted the files and reused the disk. I still have copies of the same mailboxes on the external hard drive. In this situation, it is clear that the disk was used purely for transfer. The deletion occurred not because I no longer wanted the information; I just wanted the disk space for other files. If it can be ascertained that it was a creator's regular practice to use media such as floppies primarily as a transfer device, I assert that it may be possible to assess this deleted information with less sensitivity and thus provide access to it more freely than in other circumstances.

Another viewpoint for retaining disk images is to provide technological-focused researchers an opportunity to "get under the hood" of the system in which the data was created.[xxix] Furthermore, we do not know what may take the attention of future researchers and how research may change.

How much time did I spend?

I spent two days attending the SAA Digital forensics: Advanced training course. Coupled with the reading I had already done, I felt confident to start imaging, which I did mid-September and completed my analysis and the first draft of this article by the end of November, a total of 2.5 months. I must emphasise that I did not work on this full-time; in fact most of it was done in 2-3 hour bursts in evenings and on weekends.

Breakdown of time taken on tasks

- Imaging all floppies – 1 hour, imaged 3 times – 3 hours in total
- Report generation – 2 hours as BitCurator gave me an error and I had to reboot the VM twice
- Meeting re using Droid for file identification – 2 hours
- Analysing tricky files, identifying file extensions – 3 hours
- Disk repair for bad sectors, generation and analysis – 2 hours
- "Fiddling" (e.g. mounting disks and images for comparison) – 4 hours

- Instruction sheet creation – 2 hours

I would expect that some of these times would decrease with more proficiency in manipulating the software used, looking up file signatures, and so on.

The time taken on this applied exercise not only gave me confidence to start imaging the legacy media at work, but also assisted me in estimating how much time it would take to image and complete an initial processing of the floppies (3-4 hours). In effect my estimation was pretty accurate in it taking four hours to image and appraise the work floppies. Due to the size of the hard drives the imaging of them will take longer; however, I can work on other tasks while it is underway. With the quantity of digital material I have awaiting processing, I believe that this volume is manageable and can be integrated in my workload without compromising other tasks. It would also mean that I could clear the backlog of digital material and start managing appropriately whatever is selected for retention.

Lessons learned

What is of import is not so much the type of media discussed in this article, but the opportunities afforded to work through issues that arose and options to resolve them. Since I worked on this without any immediate support, I have demonstrated that it is possible for very small institutions to embark on managing digital archives in a sustainable manner both in terms of staff time and resourcing. The greatest benefit from this exercise was the transferable skills gained.

Imaging disks and creating copies of them to appraise the contents was an efficient and satisfactory process to be able to make decisions. For those files without file extensions, viewing them through a hex editor was sufficient to gather enough information regarding the content of the files to ascertain their long-term value. The challenge comes with the file characterization that needs to occur if the file is found to

be worthy of retention, in particular completing the file validation process. Without this information, selecting appropriate preservation actions become almost impossible.

FTK Imager is more user-friendly with the hex editor included within the program. For the HFS images this meant I did not need to extract a .raw image to view them as is required in the BitCurator suite.

The strength of BitCurator is in its report generation, the contents of which assist with identifying personal information and consequently working out redaction and access conditions.

While automating processes works for those with sophisticated digital preservation systems and flush budgets, those of us who fall into the 'boutique' category will still need to depend on manual processes in the short to medium term. Ensuring that there is documentation on the processes and procedures that can be managed manually will be of invaluable assistance to small archives.

My journey through defining what is achievable for me and my institution has highlighted challenges for lone arrangers in their management of born-digital archives:

- Acquisition of technical knowledge
- Isolation
- Budgetary constraints
- Time constraints

While some of these challenges are not unique to lone arranger environments, they become more salient within this context.

The benefits of appraisal in managing my institution's legacy material make the cost of learning advanced skills worthwhile. This newfound knowledge will assist with clearing the backlog and has prepared me to work with modern acquisitions. Since the digital environment is here to stay, lone arrangers cannot be left as ostriches with their heads in the sand. If lone arranger institutions include born-digital material within their

collecting scope, then these skills are just as necessary as all the other facets of an archivist's knowledge base. On the one hand, it could appear that this is an unfair expectation for lone arrangers; however, we are still in a transition period where tools and techniques for the processing and management of born-digital collections are being explored and defined. As the archival community increasingly works with the tools and techniques applicable to the born-digital sphere and they are covered in archival beginning and continuing education programs, and as more examples of implementation are shared, the current challenges will decrease.

Due to the nature of their positions, lone arrangers work mainly in isolation unless they are part of a wider community of practice.[xxx] Overcoming this isolation requires management support for professional development opportunities and time to apply what has been learned. Furthermore, management also needs to be made aware of the complexities of digital preservation and make provision for its implementation. Acquiring requisite technical knowledge is a barrier to taking action in the born-digital sphere. Lone arrangers are confronted by all the aspects of archival administration, so it can be challenging to prioritise the learning of new skills especially when those skills are applicable to only a small proportion of the work load. In larger institutions, the tasks are divided amongst many positions (e.g. reference archivist, processing archivist, digital archivist, etc.). These institutions are now looking at ways to spread the load of digital processing to non-"digital" positions, with the digital archivists taking the lead in up-skilling colleagues.[xxxi] This method of knowledge dissemination is unavailable to lone arrangers.

Digital information requires active management to ensure its longevity. If all the data held currently on legacy media in my institution were imaged and retained, there would be an extra 500GB to store on the archives' server (which is only 232GB and 75% full). This would mean that larger on- and off-site servers would be required, as well as an increase in bandwidth available for replication to the off-site server. With limited resources at my disposal, it would be imprudent and irresponsible to keep all the images

without appraising. What I learned from my own test corpus was that material I believed I knew well still presented format identification challenges. If appraisal does not occur, then even with monitoring for fixity, the same situation will be encountered down the line with it becoming a mission to identify and access older formats, something already evidenced for files in legacy formats from the 1980s and 1990s.[xxxii] Another issue that would appear from retaining the entirety of images from hard drives, at least in my context, is the ever-increasing volume of material to manage and how that can impact on the efficient storage and retrieval of the information.[xxxiii] Espousing the retention of all images without appraising the contents replicates the issue confronting records managers where no retention and disposal schedules are applied.

A processing philosophy

Returning to Wilsey, Skirvin, Chan, & Edwards, where they articulated their position on whether to capture every bit, I was successful.[xxxiv] The challenge for the small archives is the specialized knowledge to work through file format issues. Becoming familiar with the hex editor is the first step to take. This should hopefully provide enough information for an appraisal decision to be made in the first instance. If more information is required, then recourse to one of the other methods outlined above would be the next step. While my own test collection had no content to preserve, I am unsure what my next step would be to resolve the format issue of Disk 17 if it were the 'gem' find, apart from making a call to the digital preservation community.

Deciding that too much time has been spent on troubleshooting cannot be applied wholesale. The provenance, context, and extent of the material to be imaged and processed are factors in making that decision. When more specialized or extra equipment is required for small amount of media, it may be decided that imaging will not occur.[xxxv] In fact, this statement could be broadened to include cases such as my

Disk 17 either when more specialized diagnosis is required for format identification or when the provenance of the material does not justify the expended effort.

In a general policy, it is recommended to indicate that the processing philosophy will change according to the provenance of the media being worked on. This is the method preferred by some digital archivists.[xxxvi] Again, it will be at the institution's prerogative to decide whether to deaccession or to keep successfully and unsuccessfully imaged media.[xxxvii]

This initial foray has made my provincial administration more aware of the extent of digital material currently held by and expected to arrive in the archives and what will be required to manage this material in terms of technology. Conversations to date with them indicate that our processing philosophy will change according to the importance of the contributions made by congregation members. Focusing on this philosophy will be key to ensuring the long-term preservation and access to the congregation's digital heritage within sustainable resourcing. This exercise has also finally caused my administration to consider the management of active digital records with the development of a provincial retention and disposal schedule, a much-welcomed unanticipated gain from my scholarship research.

About the author

After tertiary study at Victoria University of Wellington in French and German and teacher's college in Auckland, Elizabeth spent six years in France where she studied law in the accelerated program run by the University of Paris I – Panthéon-Sorbonne. A Master of Education followed in Melbourne, Australia. She joined the Marist Archives as a volunteer to gain practical experience while studying for a Diploma in Records and Information Management and is now the Province Archivist, a lone arranger position. She was the recipient of the 2014 Ian McLean Wards Scholarship enabling her to explore the introduction of digital preservation practices to very small archives.

Notes:

---

[i] Nancy McGovern in discussion with the author, June 1, 2015.

[ii] R. Harvey, *Digital Curation: A How-to-do-it Manual* (London, UK: Facet Publishing, 2010), [Pages 96-98].

[iii] R. Erway, *Swatting the Long Tail of Digital Media: A Call for Collaboration*, Demystifying Born-Digital (Dublin, OH: OCLC Research, 2012), [Page 4], accessed February 17, 2016, http://www.oclc.org/research/themes/research-collections/borndigital.html.

[iv] S. Welland, The Role, Impact and Development of Community Archives in New A Research Paper, 47, May 2015, accessed February 17, 2016, [page47 Zealand:], http://www.academia.edu/13398356/The_role_impact_and_development_of_community_archives_in_New_Zealand_A_research_paper._Published_May_2015.

[v] R. Chandler, "CSI Special Collections: Digital Forensics and Archives," *More Podcast Less Process*, podcast audio, October 3, 2013, accessed February 17, 2016, http://morepodcast.libsyn.com/.

[vi] R. Erway, B. Goldman, and M. McKinley, *Agreement Elements for Outsourcing Transfer of Born Digital Content*, Demystifying Born-Digital (Dublin, OH: OCLC Research, 2014), accessed February 17, 2016, http://www.oclc.org/research/themes/research-collections/borndigital.html.

[vii] deCipher Ltd. *Digital Forensics*, accessed March 30, 2016, http://www.decipher.co.nz/forensics.htm

[viii] Chandler, "CSI Special Collections: Digital."; R. Erway, *You've Got to Walk Before You can Run: First Steps for Managing Born-digital Content Received on Physical Media, Demystifying Born-Digital* (Dublin, OH: OCLC Research, 2012), [Page 3], accessed February 17, 2016, http://www.oclc.org/research/themes/research-collections/borndigital.html; K. Woods, C. Lee, and S. Garfinkel, "Extending Digital Repository Architectures to Support Disk Image Preservation and Access," in *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital*

Libraries (New York, NY: Association for Computing Machinery, 2011), [Page 58], accessed February 17, 2016, http://www.ils.unc.edu/callee/p57-woods.pdf.

[ix] M. Gengenbach, ""The Way We Do it Here": Mapping Digital Forensics Workflows in Collecting Institutions" (master's thesis, University of North Carolina at Chapel Hill, 2012), accessed February 17,

2016, http://digitalcurationexchange.org/system/files/gengenbach-forensic-workflows-2012.pdf.

[x] L. Wilsey et al., "Capturing and Processing Born-digital Files in the STOP AIDS Project Records: A Case Study," *Journal of Western Archives* 4, no. 1 (2013): [Page 11], accessed February 17, 2016, http://digitalcommons.usu.edu/westernarchives/vol4/iss1/1.

[xi] R. McNulty, "Digital Forensics in the Trenches: Floppies Gone Bad," Sulair News (blog), entry posted November 17, 2011, accessed February 17,

2016, https://lib.stanford.edu/sulairnews/digital-forensics-stanford-university-libraries/digital-forensics-trenches-floppies-gone-bad; D. Waugh, "A Dogged Pursuit: Capturing Forensic Images of 3.5" Floppy Disks," *Practical Technology for Archives*, no. 2 (2014), accessed February 17,

2016, https://practicaltechnologyforarchives.org/issue2_waugh/; Mick Crouch in discussion with the author, January 27, 2015.

[xii] Erway, *You've Got to Walk*, [Pages 3-4].

[xiii] "AFF Format Deprecated," entry posted January 15, 2014, accessed February 17, 2016, http://sourceforge.net/p/guymager/wiki/AFF%20format%20deprecated/; Kam Woods in discussion with the author, June 2, 2015.

[xiv] Dorothy Waugh in discussion with the author, May 29, 2015.

[xv] D. Waugh to BitCurator Users web forum, "Trouble Connecting to BitCurator Through a Write-Blocker," December 10, 2013, accessed February 17,

2016, https://groups.google.com/forum/#!searchin/bitcurator-users/usb/bitcurator-users/Ac_LS2AWTIU/obTR2LAfL4YJ; K. Woods to BitCurator Users web forum, "Question about Power and Write Blockers," May 25, 2015, accessed February 17,

2016, https://groups.google.com/forum/#!searchin/bitcurator-users/usb/bitcurator-

users/0fMCr61uilU/t7ee7G6QA2gJ; K. Woods to Digital Curation web forum, "Write-blocker," March 21, 2013, accessed February 17, 2016, https://groups.google.com/forum/#!searchin/digital-curation/usb/digital-curation/WLB0WLrljjU/jCMH8BPV2aAJ

xvi Waugh, "A Dogged Pursuit: Capturing,".

xvii Waugh, "A Dogged Pursuit: Capturing,".

xviii K. Woods to BitCurator Users web forum, "External 3.5 Floppy Disk Reader," October 9, 2015, accessed February 17, 2016, https://groups.google.com/forum/#!searchin/bitcurator-users/usb/bitcurator-users/0fMCr61uilU/t7ee7G6QA2gJ.

xix E. Colon-Marrero and A. Hughes, "Toni Morrison's Born-Digital Material," *Mudd Manuscript Library Blog*, entry posted August 26, 2015, accessed February 17, 2016, https://blogs.princeton.edu/mudd/2015/08/toni-morrisons-born-digital-material/.

xx A. Brown, *Practical Digital Presentation: A How-to Guide for Organizations of Any Size* (London, UK: Facet Publishing, 2013), [Page 136]; "What Is Characterization?," JHOVE2, accessed February 17, 2016, https://bitbucket.org/jhove2/main/wiki/JHOVE2_Frequently_Asked_Questions_%28FAQ%29.

xxi "Download DROID: file format identification tool," accessed February 17, 2016, http://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/

xxii "The technical registry Pronom," accessed February 17, 2016, http://apps.nationalarchives.gov.uk/PRONOM/Default.aspx

xxiii "TLM file extension," File-Extensions.org, accessed February 17, 2016, http://www.file-extensions.org/tlm-file-extension.

xxiv M. Kjörling to Digital Curation web forum, "Bad Sectors," October 20, 2015, accessed February 17, 2016, https://groups.google.com/forum/#!topic/digital-curation/-zIulSjnpho.

xxv Ross Spencer in discussion with the author, October 20, 2015.

xxvi "Photorec", accessed February 17. 2016, http://www.cgsecurity.org/wiki/PhotoRec.

xxvii P. Olsen to BitCurator Users web forum, "Floppy Drive Not Seen in Disk Management," November 20, 2015, accessed February 17, 2016, https://groups.google.com/forum/#!topic/bitcurator-users/ClXES11GbPo.

xxviii B. Goldman and T. Pyatt, "Security Without Obscurity: Managing Personally Identifiable Information in Born-Digital Archives," *Library and Archival Security* 26, nos. 1-2 (2013), accessed February 17, 2016, https://scholarsphere.psu.edu/downloads/sf268f365; S. Meister and A. Chassanoff, "Integrating Digital Forensics Techniques into Curatorial Tasks: A Case Study," IDJC 9, no. 2 (2014): [Page 13], accessed February 17, 2016, http://www.ijdc.net/index.php/ijdc/article/view/9.2.6/364.

xxix M. Kirschenbaum to BitCurator Users web forum, "We've Created a Disk Image, Now What?," August 8, 2014, accessed February 17, 2016, https://groups.google.com/forum/#!searchin/bitcurator-users/usb/bitcurator-users/1iEk7BLJB8g/vSNaomabbuQJ.

xxx Welland, *The Role, Impact and Development*, 48.

xxxi B. Gordon, "Digital Processing at the Rockefeller Archive Center," *bloggERS!*, entry posted April 5, 2016, accessed April 6, 2016, https://saaers.wordpress.com/2016/04/05/digital-processing-at-the-rockefeller-archive-center/.

xxxii C. Rusbridge, "The PowerPoint 4.0 Adventure: What Did I Learn?," *Unsustainable Ideas* (blog), entry posted October 15, 2012, accessed April 6, 2016, https://unsustainableideas.wordpress.com/2012/10/15/ppt-4-adventure-learning/.

xxxiii C. Prom to Digital Curation web forum, "Why Do We Save Disk Images?," March 6, 2014, accessed April 24, 2016, https://groups.google.com/forum/#!topic/digital-curation/vxbd7t-mAfo.

xxxiv Wilsey et al., "Capturing and Processing Born-digital," [Page 11].

[xxxv] Dorothy Waugh in discussion with the author, May 29, 2015

[xxxvi] Wendy Hagenmaier and Kathryn Michaels in discussion with the author, May 28, 2015.

[xxxvii] J. Durno to Digital Curation web forum, "Digital Media Items Post-imaging," April 23, 2016, accessed April 24, 2016, https://groups.google.com/forum/#!topic/digital-curation/37fOLcXjmc8.