Notes and Communications

Five Hundred 5.25-Inch Discs and One (Finicky) Machine: A Report on a Legacy E-Records Pilot Project at the Archives of Ontario

CHARLES LEVI

RÉSUMÉ Il existe un vaste corpus de textes théoriques portant sur le sujet des documents numériques, de leur authenticité et des problèmes liés à leur gestion. Cependant, il existe une lacune au niveau des compte rendus du traitement des documents numériques, surtout en ce qui a trait aux documents numériques patrimoniaux sur disquettes. En 2008, les Archives de l'Ontario ont lancé un projet-pilote sur les documents numériques afin d'analyser les problèmes engendrés par ces disquettes. Jusqu'à maintenant, on a évalué plus de cinq cents disquettes 5,25 pouces, six cents disquettes 3,5 pouces, cinquante cédéroms, deux lecteurs ZIP et plusieurs autres médias. Bien qu'on ait connu un succès considérable avec l'utilisation des lecteurs de disquettes 3,5 pouces externes, le défi de maintenir un lecteur de disquettes 5,25 pouces fonctionnel se poursuit. Ce texte rend compte de : 1) le progrès du projet des documents numériques; 2) les problèmes reliés à la lecture et à l'évaluation des fichiers électroniques, y compris ceux qui ne sont plus supportés ou qui sont de format obsolète; 3) l'investigation légale poussée des logiciels; 4) les nouveaux développements possibles dans le domaine du matériel informatique. L'accent est placé sur la récupération de l'information plutôt que sur la préservation des médias archaïques.

ABSTRACT There is a vast amount of theoretical literature on the subject of electronic records, their authenticity, and problems with their management. There is, however, a lack of practical reporting on the processing of electronic materials, especially legacy e-records on floppy disk. In 2008 the Archives of Ontario initiated an e-records pilot project to analyze the issues arising from these floppy disks. To date, over five hundred 5.25-inch floppy disks, six hundred 3.5-inch disks, fifty CD-ROMs, two zip-drives, and a variety of other media have been appraised. Although there has been considerable success in the use of 3.5-inch external disk drives, the struggle to maintain a usable 5.25-inch disk reader has been ongoing. This paper reports on: 1) the progress of the e-records project; 2) the issues involved in reading and assessing computer files, including those of currently unsupported and obsolete formats; 3) advanced software forensics; and 4) possible future hardware developments. The focus is on the recovery of information as opposed to the preservation of archaic media.

ARCHIVARIA 72 (Fall 2011): 239–246

In 2008, the Archives of Ontario began its second legacy electronic media pilot project. This project has now moved beyond the pilot stage, and its activities and findings should be of use to any archivist dealing with such media, especially floppy disks of the 3.5- and 5.25-inch variety. This paper will discuss how electronic media arrives at the Archives, our methodology for evaluating this material, the machinery we have used to do this (and possible alternatives), and some issues about migration and reformatting. The focus will be on recovery and preservation of information as opposed to the preservation of physical containers for that information.

The first "obsolete media" project at the Archives of Ontario took place between 1999 and 2004, but its existence was not widely known when the second project began in 2008. The first project cost the Archives $6,000, "hundreds" of hours of staff time, and dealt with 871 media in eleven different types. To date, the second project has cost essentially nothing, but also has used hundreds of hours of staff time and has dealt with more than fifteen hundred media in eight different types (see Appendix A). We do not know, however, what percentage this represents of the total extent of our legacy electronic media. Although the first project's mandate was to find all of the legacy records then in our holdings, media continues to surface in unexpected places.

Our legacy records come in boxes. The Archives of Ontario annually receives thousands of boxes of government records that have reached the inactive stage of their life cycle. Although government record schedules and best practices for records management dictate that electronic media *should not* be sent to the record centre, these rules are not always followed to the letter. Every box which is opened carries the possibility of containing a floppy disk, CD-ROM, or other media item which, according to the schedule, should not be there. More predictable are electronic legacy media in private donations. Our assessment process for these donations normally identifies the extent and nature of this media. As has been discussed in previous archival conferences, "Digital Lives" are becoming more evident, and there is extensive documentation on this – and how to deal with it – available on the website of the British Library.[1] The discussion which follows, however, does not discuss the particulars of private records.

Through the life of the current pilot project from 2008 to 2010, 389 5.25-inch disks and 938 3.5-inch disks have been located either within new accruals of government records or during backlog listing projects. The remaining quantity of disks currently in our established holdings, as well as those that

1 See Jeremy Leighton John, "Adapting Existing Technologies for Digitally Archiving Personal Lives: Digital Forensics, Ancestral Computing, and Evolutionary Perspectives and Tools," iPres 2008 Conference Paper, http://www.bl.uk/ipres2008/presentations_day1/09_John.pdf (accessed on 8 June 2011); John A. Blythe, "Digital Dixie: Processing Born Digital Materials in the Southern Historical Collection" (Master's Thesis, School of Information and Library Science, University of North Carolina at Chapel Hill, July 2009).

lurk in record centre boxes scheduled for transfer over the next thirty years, is unknown. They do not appear prominently in either file or transfer lists from government ministries and agencies. Recently I had the opportunity to create a complete file listing and description for the Sewell Commission (which dealt with planning in Ontario in the 1990s). The Sewell team were enthusiastic users of WordPerfect™ 5.1, and left behind thirteen 5.25-inch disks of files (as well as memoranda describing their training on the use of the software). These were not listed in the collection files, or in the extensive documentation that the Commission provided to the Archives when it transferred their records. But they existed.

We are obliged, as archivists, to appraise this and other such material upon its near-random appearance within government records. The Archives rejected outright the possible outcome that since this material was not scheduled to come to us, we did not need to "see" it. But with legacy material we have taken an appraisal approach that largely rejects the strategies for "born digital" records advocated elsewhere, such as the InterPARES project.[2] The Archives has done so with the understanding that the definition of "born digital" in the current context is not the definition that would have been used in a government office in the 1980s and 1990s. This understanding has evolved as we continue to analyze the context in which these records appear, and the true instability of magnetic media, especially floppy disks of the 5.25- and 3.5-inch variety. As mentioned above, the focus has been on information preservation, not on media preservation. This has been the only viable appraisal criterion.

Disks are found in boxes. They are also often found in or near files that contain printouts of computer-created data, such as word processing files (including correspondence and forms), simple spreadsheets, charts and graphs, presentation slides, and other such applications. In many cases the documents are labelled the same way they are on the disks (complete with file name and extension). Although the use of computerized platforms to create this data implies the records are "born digital," the aggregate experience of appraising these materials strongly implies that those who used the tools had no intention of

2 This is an inference, since the Glossaries prepared by InterPARES 2 and InterPARES 3 define "digital" and "born digital" with reference to machines and not humans. InterPARES 2 defines "digital" as "the representation of an object or physical process through discrete, binary values. In contrast to an analogue representation of an object or physical process, a digitally-encoded representation does not resemble the original," http://www.interpares.org/ip2/ip2_terminology_db.cfm?letter=b&term=211 (accessed 21 July 2011). InterPARES 3 defines digital as "The representation of a physical process through discrete, binary values," http://www.interpares.org/ip3/ip3_terminology_db.cfm?letter=d&term=211 (accessed July 21, 2011) and "born digital" as "Originally generated in digital form," http://www.interpares.org/ip3/ip3_terminology_db.cfm?letter=b&term=679 (accessed 21 July 2011). InterPARES has mostly been focused on records created within document management systems and databases – the extension of their theories to floppy disks has not been asserted to date.

leaving those items in digital format. Much like the "old-timers" who print and file their email or PowerPoint™ slides from presentations attended, late-twentieth-century users considered the computer a transitory tool designed to move information from mind to paper. WordPerfect™ 5.1 was thus merely a method of digitally expressing notions born in the human brain – a slightly more sophisticated Dictaphone.[3]

Our acceptance of the principle of digital expression as a means toward a paper end product, influences our appraisal processes. Thus, if the information is already available in printed form, we do not preserve the electronic version. If one out of twenty documents on a disk was not printed, we print the one document. If a reasonably sized spreadsheet can be accessed and printed, it also gets printed. Both "Cull" and "Print and Cull" represent the vast majority of our appraisal decisions for legacy media. The only electronic documents kept are those that are "too big or too complex" to print. This represents less than one percent of all legacy media discovered (See Appendix A).[4]

To get to this stage, however, you have to be able to read the information in the media format on which it arrives. It is this latter task that has become the most complex part of the process. In most cases, we have benefited from the tendency of government agencies to use common applications that are still readable. Microsoft Word™ files on 3.5-inch disks pose little problem – especially given the widespread availability of portable 3.5-inch drives with USB capability.[5] Our computers are no longer equipped with floppy drives; we do, however, have six 3.5-inch portable drives that can be signed out and used as necessary. It seems safe to assume that these drives will still be commercially available in the near future (although it is unclear as to whether they will be compatible with future operating systems).

More troublesome is the issue of the 5.25-inch disk. As Jeremy John of the British Library declared in 2008, "the use of ancestral computer technology for digital capture is unavoidable at present."[6] And he is right. Finding a system that could read a 5.25-inch disk was possible with the first project in 1999, which listed 5.25-inch disks as one of the three media that the Archives had the capac-

3 The mid-1980s situation has been partially expressed by a WordPerfect™ employee: "Law firms were especially loyal to the old machines, because many lawyers were not ready to have computers on their desks. They liked the old way of dictating their documents to their secretaries and having final versions produced by the firm word processing center." See W.E. Peterson, *Almost Perfect*, chapter 7 (1998), http://www.wordplace.com/ap/ap_chap07.shtml (accessed on 8 June 2011). The next step – to eliminate the secretaries – would follow.

4 Most of our appraisal decisions have dealt with small groups of files clearly reproduced in the paper record; the decisions would be quite different if a significant amount of scientific data not committed to paper was located. It is rare, however, for such information to be found in its preservation version on a floppy disk.

5 This "widespread availability" might not last; Lenovo discontinued their production of these devices in late-2009 (or early-2010).

6 John, p. 4.

ity to process, and noted with concern that there were "no facilities within the government" to read eight other types of media. In 2008, this list had jumped to nine. The Archives used a machine provided by a staff member's brother-in-law. It came with no manual, no warranties, and no support. When it broke down, the three most qualified computer-hobbyists on the Archives staff could not repair it. Nor could a machine be found commercially to replace it. Or for free. This is not meant ironically; an organization in Toronto called FreeGeek rebuilds and recycles old computers in return for goodwill and supplies. They, however, could not produce a machine with a working 5.25-inch drive. Within a ten year period, 5.25-inch drives had gone from reasonably standard to completely absent.

In the process of discussing these issues, we fortuitously discovered a second government agency that also had a 5.25-inch drive "problem"; it too had a staff member prepared to try to fix the problem. He found some 5.25-inch drives in a Toronto basement, and constructed a machine capable of both reading 5.25- and 3.5-inch disks in a Microsoft Windows™ environment and copying them to either a CD-ROM or an external hard drive using USB 1.0 technology. The external case of this machine, however, was an early-2000 design – current government computers do not have the proper conversion cables, making the machine a "one-off." It most certainly has a fixed lifespan, but it did come equipped with six spare 5.25-inch drives, which we hope we can use as needed.

In building this machine, the government recreated the efforts of the Prometheus Project, an initiative of the National Library of Australia and a project unknown to the Archives of Ontario. The Prometheus Project created mobile "mini-jukeboxes," with varying drives that could be moved among workstations. However, as I read the current specifications for Prometheus, it does not include 5.25-inch drives.[7]

We do not know when we will come across our last 5.25-inch disk. They were supplanted by 3.5-inch disks in 1987, but were still used into the 1990s. With a possible thirty-year retention schedule by some agencies, we may have 5.25-inch disks arriving for the next ten years. Will we have the capacity to read these disks? Will our machine last that long? We do not know.

The Archives of Ontario is also monitoring the progress of a European company called Kryoflux. This company is currently producing an interface that allows a 5.25-inch disk drive to be connected to a USB 2.0 port, which reads and verifies the bitstreams produced on the disk, and can also convert these to disk image files. The company created this product to support their key interest: to verify the authenticity of heritage computer video games in order to preserve

7 National Library of Australia, "Prometheus Installation Guide," 25 November 1998, http://superb-sea2.dl.sourceforge.net/project/prometheus-digi/Documentation/Prometheus_Component_Installation_Guide_SF-v1.pdf (accessed on 8 June 2011).

them in their original form. Unfortunately, in its present format, Kryoflux is not user-friendly. The British Library was an early adopter, and I had the good fortune of having the product demonstrated to me by the Library staff in 2010. I still have no idea exactly what it does or whether it is useful, and I am still waiting for an understandable description of its capability to be written in archival-friendly language by those who have become "early adopters" of the product. The commercial license for Kryoflux (between CDN $4,000 and $8,000) is a reasonable price, considering the development costs of the product, but may be beyond the reach of most institutions. And, maddeningly, the product is still hardware dependent – if you do not have a 5.25-inch drive of your own, you cannot use the interface.[8] Archival repositories need a commercial operator/developer willing to return to the business of producing 5.25-inch drives. Such an endeavour would, however, require a profitable market, which is unlikely. Failing that, a consortium of archives could sponsor an archival 5.25-inch construction project. The schematics and materials must still exist somewhere.

The last major issue this paper will cover is file conversion, reformatting, and migration. This is especially tricky for the 5.25-inch computing universe. Younger archivists are often surprised to discover that there was a time when standardized file extensions did not exist. Until the rise of Microsoft Windows™ (and, indeed for some time after), people routinely used whatever naming convention they understood – and rarely left explanatory notes. They could, unwittingly, use extensions that would be later assigned to proprietary software. Sometimes half the challenge is figuring out exactly what type of file you are dealing with. Even the "standardized" extensions of the 1990s were used by multiple programs (e.g., the extension .cht – used by both Borland's DBASE and Harvard Graphics) as well as several other niche applications.[9] There are resources on the World Wide Web that can help you determine what sort of files you might encounter – and even some quick conversion hints. This is how I determined that an old Lotus 1-2-3™ spreadsheet can become an Excel™ spreadsheet simply by changing the file extension from .wk1 to .xls – data loss is minimal.[10] Similarly, Quattro Pro™ can be converted through an existing patch to Excel™ – at least for the near future.[11]

Some conversion methods require more ingenuity. Older versions of Adobe Acrobat™, such as version 2.1 released in 1995, seemed to require a machine

8 For more on Kryoflux, visit http://www.kryoflux.com (accessed on 7 July 2011). Their specifications fluctuate daily as they are still working on their product.

9 See http://filext.com/file-extension/CHT (accessed on 8 June 2011). .cht might also be an ICQ-saved chat session or a chart created by "My Health Software." Neither of these has surfaced yet at the Archives of Ontario.

10 See comments by "Mindcore," http://www.thecomputermechanics.com/forums/showthread.php?44502-wk1-to-xls (accessed on 20 June 2001).

11 See http://office.microsoft.com/en-us/excel-help/opening-quattro-pro-files-in-excel-HA001044873.aspx (accessed on 20 June 2011).

that could run Windows 95™, as well as some way of compressing and uncompressing the files to transfer them between computers. Unlocking and reading such files is possible; it took me four computers to do it, at which point I discovered the material was duplicate and could be culled. I have since discovered that the file management system used by the Archives of Ontario (Hewlett Packard's TRIM™) has the capacity to read Adobe™ 2.1 files. It could not, however, have run the executable file that was required to extract the Adobe. pdf data. TRIM™, as well as other file management programs, has a remarkable but ill-advertised capacity to open files from long-lost applications such as Paradox™, RBase™, Multimate™, and Wordstar™ – if you can get them off the disk onto CD-ROM.

And if you can identify them. This is the promise of a few other applications we are testing, such as JHOVE (the JSTORE/Harvard Object Validation Environment – based at Harvard University Library),[12] and DROID (The Digital Record Object IDentification tool offered as a free service from the UK National Archives).[13] Although these applications are beyond the scope of the present paper, suffice it to say that both are tools designed to use computer forensic techniques to attempt to identify the file format of unknown files by exploiting common digital structures found in known examples. Neither application is in the standard archival workflow yet; we are using them to deal with poorly documented electronic material that was previously processed, but not for new arrivals. Another project currently underway aims to bring changes to our current accessioning process for electronic records, so as to allow for more detailed technical examination of the records, including verifying file formats and possibly, normalization. However, in the future a seamless link between data recovery on our remanufactured machine, file analysis through these programs, and ingest to a Trusted Data Repository (in the development stages) will be possible. If this is combined with robust hardware support for antiquated disk drives, the appraisal of electronic records will go from chaotic nightmare to placidity.

In conclusion, what have we learned from the e-pilot project at the Archives of Ontario? First, we now know that the advice that is sometimes given – that e-records analysis requires outsourcing to private companies – is erroneous. The technological supports exist for on-site appraisal and analysis, if you have the right machines. Second, the need for getting the right machines together and working is immediate, and the clock is against us. Computer rebuilding is a complicated matter and the skill sets that support it are not always easy to find. Third, the Archives of Ontario should not be a museum of obsolete technology; that said, some embracing of the ancient machine will be required for at least

12 See http://hul.harvard.edu/jhove/ (accessed on 10 June 2011).

13 See http://www.nationalarchives.gov.uk/information-management/projects-and-work/dc-file-profiling-tool.htm (accessed on 10 June 2011).

the next decade – by which time the storage devices we now think of as common might also be heading toward extinction. As archivists, we should always be one generation behind – difficult in a government that tends to upgrade all machines at once. Fourth, the number of employees who have first-hand experience with computers of the 1980s and early-1990s is dwindling, both in the working world and in the world of archives. Finally, the scattered knowledge on the World Wide Web regarding obsolete machinery and methods of retrieval is an inefficient base on which to support the appraisal of legacy e-records. Some sort of manual or guide to the world of the 5.25- and 3.5-inch floppy disk is badly needed for the edification of working archivists.

When the Archives of Ontario started its current pilot in 2008, we had five hundred 5.25-inch disks and a finicky machine. Now, in 2011, we have an unknown number of 5.25-inch disks and a slightly less temperamental computer on which to process them. We have not solved the legacy records problem, but we have – through a philosophy of information preservation and a survey of available tools – lessened its impact on our appraisal workflow.

Appendix A
Some Statistics on the Pilot Project 2008– 2010

Total employees recording appraisals	17	
Total number of appraisals	289	
Average appraisal time	48	minutes
Total staff time spent on appraisals	32	days
Appraisal decisions		
Cull	182	
Print and Cull	65	
Defer/Retain	31	[items too voluminous or in unreadable formats]
Migrate	9	
Not recorded	3	
Media types		
3.5-inch Floppy Disk	938	
5.25-inch Floppy Disk	389	
CD-ROM	114	
½-inch data reel	55	
8-inch Floppy Disk	5	
IBM magnetic cards	3	
ZIP disks	2	
Micom Disk	1	
Unspecified	30	