Volume 42 | Number 1 Article 6

7-1-2014

The Tail Wagging the Dog: Challenges of Working with Obsolete Computer Media

Ben Goldman
Pennsylvania State University, bmg17@psu.edu

Contact Joanne Kaczmarek at jkaczmar@illinois.edu if you would like to guest author an Electronic Currents column or share a good idea.

The Tail Wagging the Dog: Challenges of Working with Obsolete Computer Media

By Ben Goldman, Pennsylvania State University

Thanks to SAA's Jump In Initiative[1] and the publication of an OCLC research report,[2] archivists from a variety of institutions are now surveying collections for fugitive computer media and implementing strategies for recovering data from such media. But as archivists dig into legacy holdings, they are uncovering the "long tail" of computer media: more exotic formats beyond typical floppy disks and optical media. Getting to the hidden content contained in these formats requires access to elusive equipment and strategies that may seem too prohibitive to be worth any one institution's resources. This paper recalls the trials and tribulations of a project by Penn State Special Collections to recover content from old computer media. It closes by pondering the possibilities of working collaboratively across institutions to recover legacy data files.

Setting the Stage

While preparing a collection for processing, Penn State staff members turned up 27 three-inch, double-density Amstrad[3] computer disks in the literary manuscripts of a modern English author. The disk labels indicated that many contained writings but did not indicate whether the writings replicated materials found in the analog part of the collection. A quick Internet search determined the disks were used on an Amstrad CPC or PCW computer, both briefly popular in the United Kingdom and Europe during the 1980s. Amstrad computers used the CP/M operating system and were popular for word processing using LocoScript software.

Some Amstrad machines are affordably available through websites such as eBay. But since vintage machines often do not function properly and because we couldn't be certain which model would work with the disks we had, we didn't purchase one. The author's analog papers were reviewed for clues about the machine used to create the files, but only 1993 correspondence indicating the decision to purchase a new Macintosh was found. We contacted the author, who couldn't recall any details of the Amstrad or the exact years of its use. We pondered the wisdom of purchasing obsolete equipment that might never be used locally again. Because we could not answer with certainty whether the data contained on the disks merited the investment of time and archival resources, we chose not to purchase.

If not for an OCLC Research project[4] exploring the feasibility of outsourcing recovery of fugitive media, we might have decided the effort to proceed was too great. With the help of an internal research grant, we decided to use these 27 Amstrad disks as a way of contributing to the ongoing investigations led by Ricky Erway at OCLC Research.[5]

Project Details

Using the protocols piloted by OCLC Research, which included the development of a template agreement detailing responsibilities and roles, we agreed to send the disks to an American vendor. We also outlined our desired deliverables: a set of 54 disk images (two for every double-sided disk) in a standard image format, named according to a convention documented in a Google spreadsheet, and including a checksum for each image. Unfortunately, the vendor could not read the disks due to the limited availability of Amstrad hardware in the United States. We found a vendor in the United Kingdom who had experience with Amstrad machines, and we provided two sample disks.

Despite the agreement with the vendor, ultimately *none* of our requirements were met. Reading the disks proved challenging and took months. Commercial Amstrad machines include the CPC series, popular for gaming capabilities, and the PCW series, most commonly used for word processing. The vendor concluded the disks were used with the PCW series, probably a PCW 9512, but he did not have a functioning model. The vendor found a computer enthusiast in Cornwall who verified that he could read and recover the disks. We then shipped the remaining disks to the computer enthusiast.

We eventually received data, though not in the manner we had hoped. Disk images were produced, but the images were in .DSK format, imposed by the native operating system of the Amstrad machines. We could do little with this format beyond rendering it in an Amstrad emulator; neither of the two forensics tools adopted at Penn State (BitCurator and FTK Imager[6]) supports the CP/M operating system, and there is no known way to migrate the .DSK file to another disk image format. Additionally, the vendor provided three versions of every file found on the disks: RTF, Word 95, and Word 2.[7] These file formats were chosen to account for various levels of data loss when migrating out of the

LocoScript software. Line spacing, tab stops, and font sizes were three particular issues noted. In sum, our recovery efforts ended with three versions of every unique file (861 in total) rendered to more modern but not fully contemporary formats, each with its own particular brand of data loss, in addition to a master disk image that was unusable outside the native operating system.

Lessons Learned

This experience provided some useful lessons. Outsourcing tasks to recover data from obsolete disks needs to be accompanied by clearly defined expectations, but technology sometimes defies curatorial intentions. Unforeseen challenges might require different strategies or solutions that contradict best practice around digital preservation. So while our relationships with vendors must be structured through written agreements, we also need to be flexible when technology hurdles get in the way of meeting requirements.

It was somewhat dizzying to discover that we have obsolete disk image formats. This begs numerous questions. Will the profession have to worry about format obsolescence with disk images too? Will archivists be forward-migrating images en masse in 20 years? What other challenges have we not anticipated? Would we have pursued this project if it hadn't been funded by a grant? It's hard to say.

Despite not having the archival requirements met in the recovery of the data, we still recovered data. We are processing the digital files in the collection and gaining much needed experience with hybrid collections. Disk images only usable in the native operating system present us the opportunity to explore emulation tools for researchers. Perhaps there are other archival repositories with significant holdings on Amstrad disks that will benefit from our experimentation. More questions come to mind. Is outsourcing media recovery a sustainable strategy for archives? Perhaps the price of outsourcing is too steep, especially when the disk content cannot be appraised in advance.

Shortening the Long Tail

The Penn State Special Collections experience with Amstrad disks has us thinking about building a collaborative framework for accomplishing this work within the profession. Many institutions now have born-digital archives programs outfitted with varying levels of technical infrastructure and an array of media types. Our workflows may differ, but they are all informed by the same underlying assumptions about digital preservation, which provides an excellent baseline for developing requirements for outsourcing among institutions. But we lack documentation on our practices, workflows, and the equipment procured to accomplish this work. A good start in this direction might simply be establishing a kind of formal registry where archivists can document the hardware and software infrastructure they have in place and to which we could all refer when new local challenges arise. Ultimately, beyond documentation and a formal registry, I imagine we will need a cooperative approach to recovering legacy data files. We don't want the long tail of computer media wagging the archival dog. The best chance we have of dealing with the variety of unusual media found in archives will require some intentional coordination of our efforts.

Notes

1. Society of American Archivists' Manuscript Repositories Section Jump In Initiative encourages archivists to just "jump in" to managing born-digital content by asking them to take first steps and submit a short report about their experiences. Society of American Archivists, accessed May 5, 2014, www2.archivists.org/groups/manuscript-repositories-section/jump-in-initiative.
2. Ricky Erway, "You've Got to Walk before You Can Run: First Steps for Managing Born-Digital Content Received on Physical Media," OCLC Research, August 2012, accessed May 5, 2014, www.oclc.org/content/dam/research/publications/library/2012/2012-06.pdf.
3. Amstrad is a British electronics company founded in 1968. *Wikipedia*, s.v. "Amstrad," accessed May 5, 2014, en.wikipedia.org/wiki/Amstrad.
4. Ricky Erway, "Swatting the Long Tail of Digital Media: A Call for Collaboration," OCLC Research, 2012, accessed May 5, 2014, www.oclc.org/research/publications/library/2012/2012-08.pdf.
5. OCLC Research piloted a test data recovery project that helped establish a baseline for costs ($40 per disk) and proposed protocols, while also informing the development of a template agreement (forthcoming) for archivists to use when working with vendors.
6. BitCurator is a suite of freely available open source forensics tools repackaged for use in archives, accessed May 5, 2014, www.bitcurator.net. FTK Imager is a free version of the commercial digital forensics software from Access Data, accessed May 5, 2014, www.accessdata.com.
7. Because Word 2 files have trouble rendering in some updates to Word 2003, the vendor also supplied workaround scripts so that the files would display correctly.