FixityBerry: Environmentally Sustainable Digital Preservation for Very Low Resourced Cultural Heritage Institutions

Anthony Cocciolo, Pratt Institute School of Information and Library Science

Abstract
Whereas large cultural heritage institutions have made significant headway in providing digital preservation for archival assets—such as by setting-up geographically redundant digital repositories—medium and small institutions have struggled to meet minimum digital preservation standards. This project will explore one option for enhancing the digital preservation capacity for very low-resourced environments. FixityBerry is a project which connects consumer-grade USB hard disks to the $35 Raspberry Pi computer, which checks file fixity weekly and powers down when checking is complete. This poster will report out on an eight-month pilot of using FixityBerry to monitor the digital assets from several small cultural heritage institutions.
Keywords: digital preservation, cultural heritage, system evaluation
Citation: Cocciolo, A. (2015). FixityBerry: Environmentally Sustainable Digital Preservation for Very Low Resourced Cultural Heritage Institutions. In *iConference 2015 Proceedings*.

Contact: acocciol@pratt.edu

1 Introduction

Whereas large cultural heritage institutions have made significant headway in providing digital preservation for archival assets—such as by setting-up geographically dispersed digital repositories—medium and small institutions have struggled to meet minimum digital preservation standards. A simple but demonstrative example of this problem is that digital preservation standards require that digital assets, when accessioned into an archive, be incorporated into an enterprise storage system (e.g, NDSA, 2013). Many small cultural heritage institutions don't have such systems and instead rely on consumer hard drives available for purchase at retailers like, or rely on free storage from sites such as the Internet Archive. The discrepancy between digital preservation standards and the ability for small and medium-sized cultural heritage institutions to meet those standards has prompted researchers to investigate how this chasm can be closed. This is best highlighted by the U.S. IMLS-funded POWRR project (Preserving [Digital] Objects with Restricted Resources), which is studying strategies for medium and small institutions to provide long-term preservation to digital assets (Rinehart & Prud'homme, 2014).

Figure 1. FixityBerry, which includes a Raspberry Pi connected to USB hard drives.

In a similar concern, this project is motivated by the question: *How can very low resourced cultural heritage institutions provide long-term preservation for their digital archival assets?*

This project will explore one option for enhancing the digital preservation capacity for low-resourced environments. FixityBerry is a project which connects consumer-grade USB hard disks to the $35 Raspberry Pi computer, which checks file fixity weekly, emails the archivist a report, and powers down when checking is complete (unit shown in Figure 1). This poster will report out on an eight-month pilot of using FixityBerry to monitor the digital assets from several small cultural heritage institutions, including the Lesbian Herstory Archives in Brooklyn, NY, the Archives of the Center for Puerto Rican Studies, Hunter College, City University of New York, among others.

2 Literature Review

Very low-resource environments can include a wide-variety of cultural heritage institutions, such as local or municipal historical societies, ethnic archives, LGBT archives, among others. In these contexts, if digital material is included in a donation, such as on a USB hard disk, the original media will often be

placed in acid-free boxes or on a shelf and the media not imaged or incorporated in a storage system, as is the best practice in handling removable media (Barrera-Gomez & Erway, 2013). Best practice further dictates that file fixity is monitored regularly, such as performing a checksum and verifying it against original accession records, which is difficult to do when media are dispersed across boxes (CRL/OCLC, 2007). Hard drives laying dormant in boxes can pose additional problems. For example, Hughes and Murray (2005) note that prolonged un-operating hard disks can suffer from stiction, which is where the disk head cannot move to the appropriate position on the disk during a read operation because of lubrication failure. Further, old disks can suffer from corrosion (Hughes & Murray, 2004). Thus, low-resourced cultural heritage institutions can struggle to meet basic standards for digital preservation. A high standard for digital preservation is captured in Trusted Repositories Audit and Certification checklist (CRL/OCLC, 2007). A simplified—albeit less thorough standard—is provided by the National Digital Stewardship Alliance's Levels of Digital Preservation.

Research has demonstrated that bit rot, bit flipping, or silent data corruptions occur in all storage systems, and that a way to demonstrate that this has occurred is through file fixity checks (Rosenthal, 2010; Bairavasundaram et al., 2008). For example, Bairavasundaram et al. (2008) studied 1.5×10^6 hard drives over a period of 41 months, and discovered 4×10^5 silent data corruption incidents. Similarly, Rosenthal (2010) notes a study of data storage at CERN (the European Organization for Nuclear Research) from Kelemen (2007), who discovered that 1.2×10^{-9} of the data written to CERN's storage was permanently corrupted within six months. In libraries and archives, bit rot has proven to be a persistent problem. In imaging media from the Stop AIDS Project that was donated to the Special Collections of Stanford University, staff there found that only 8% of CDs were successfully imaged, 60% of 3.5 floppy disks, and 96% of zip disks, all of which were the result of some form of data corruption (Wilsey, Skirvin, Chan & Edwards, 2013).

Fortunately, Baker et al. (2006) find that consumer grade hard drives are not necessarily less reliable than enterprise hard drives, although they are far less expensive. Thus, it is reasonable to believe that using consumer grade hard disks, combined with secondary and tertiary copies with fixity monitoring, can enhance digital preservation for low-resourced environments.

3 System Design

FixityBerry employs the $35 Raspberry Pi computer, with a USB splitter ($10), 8 GB memory card ($15), and 3 USB hard drives (500 GB Western Digital, 1 TB Hitachi, 4 TB Western Digital). Any combination of USB drives can be used (e.g., hard disks, thumb drives, etc.); these were used because the researcher already possessed them. The FixityBerry script, which is available from download at GitHub, uses PHP to run the MD5 checksum algorithm, where the checksum values are stored in the Pi's MySQL database. The FixityBerry PHP script monitors directories for new additions and changes (e.g., file removal, new file added), and sends alerts after running via email to the recipient, and indicates things like checksum validation errors. Minor additions and changes to the default Raspbian Linux Operating system are necessary and included with the setup script, such as auto-mounting USB drives as read-only drives, installing packages to read Macintosh and Windows file systems, and changes to startup and shutdown procedures. The setup documentation is included on FixityBerry's GitHub repository page.[1]

To make the system as environmentally sustainable as possible, the hard drives and Raspberry Pi are powered off when fixity checks are not in progress, thus making minimal electrical demands. To achieve this, the entire unit (Raspberry Pi and USB Hard drives) is connected to a power strip, which is connected to a Stanley Digital Power Socket timer ($10), which shuts down the unit when fixity checking is not in progress. Since the first run of FixtyBerry took approximately 36 hours to complete, the researcher set the timer to activate the power strip on Thursday afternoon, and power-off on Saturday afternoon, for a total of 48 hours. Thus, FixityBerry would run during this time, gracefully shutdown the Raspberry Pi when complete, and the power timer would terminate the power to the computer and drives. In sum, the fixity checks get run weekly, ensuring the files are checked for fixity and that the hard drives are not subject to degradation from non-use. If a file is found to fail checksum validation, it can be restored from a secondary backup copy.

[1] https://github.com/acocciolo/fixityberry

The total cost of all components is $70 (excluding hard disks). As new hard disks get loaded up with content for checking, they can be simply be plugged-into the unit without any additional configuration.

4 Evaluation

FixityBerry is being run over the course of eights months (July 2014 to March 2015) on collections held by the researcher for the following small archival institutions shown in Table 1. As of October 1, 2014, the unit has performed as designed: weekly emails were sent from the unit resulting in 13 scans, and no checksum errors were found during that time. By March 2015, the eight-month evaluation will be complete and full results will be reported-out at the *iConference 2015*. Particularly noteworthy finding will include any checksum validation errors found or other system failures, such as failures of FixityBerry to run or email the results of the scans.

Table 1. Digital files checked using FixityBerry

Archive	*Holdings*	*Number of Files*	*Total GB*
Lesbian Herstory Archives	digitized oral histories (audio and video)	7,362	1,228.8 GB
Archives of the Center for Puerto Rican Studies (Hunter College/CUNY)	digitized oral histories (audio)	6,488	136 GB
Interviews with dancers by dance critic Barbara Newman (personal collection)	digitized interviews (audio)	16,271	351 GB
Backup of *German Traces NYC* website (educational project from Pratt Institute)	Backup of video, audio, and website files	4,288	80 GB
Total:		34,409	1,795.8 GB (1.8 TB)

5 Conclusion

In conclusion, FixityBerry provides an environmentally sustainable, low-cost method to monitor digital files for long-term preservation. This was evaluated by connecting 1.8 TB of digital assets from small cultural heritage institutions to FixityBerry, which are scanned weekly for fixity. The initial three-month evaluation indicates that the setup achieves its goals by allowing for a low cost method for monitoring file fixity and keeping hard drives from prolonged non-use that can lead to stiction. Individuals and institutions interested in reproducing their own FixityBerry setup can learn how from the project's website on GitHub.

References

Bairavasundaram, L. N., Arpaci-Dusseau, A. C., Arpaci-Dusseau, R. H., Goodson, G. R. & Schroeder, B. (2008). An analysis of data corruption in the storage stack. *ACM Transactions on Storage, 4*(3). doi: 10.1145/1416944.1416947

Baker, M., Shah, M., Rosenthal, D. S. H., Rossopoulos, M., Maniatis, P., Giuli, T., Bungale, P. (2006). A fresh look at the reliability of long-term digital storage. *Proceedings of the EumSys'06, April 18-21, 2006, Leuven, Belgium*. New York: ACM, pp. 221-234. doi: 10.1145/1217935.1217957

Barrera-Gomez, J. & Erway, R. (2013). *Walk This Way: Detailed Steps for Transferring Born-Digital Content form Media You Can Read In-house*. Dublin, OH: OCLC Research. Retrieved from http://oclc.org/content/dam/research/publications/library/2013/2013-02.pdf

Center for Research Libraries/Online Computer Library Center. (2007). Trustworthy repositories audit & certification: criteria and checklist. Retrieved from http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf

Hughes, G. F. & Murray, J. F. Reliability and security of RAID storage systems and D2D archives using SATA disk drives. *ACM Transactions on Storage, 1*(1), 95-107. doi: 10.1145/1044956.1044961

Kelemen, P. (2007). *Silent Corruptions.* In *8th Annual Workshop on Linux Clusters for Super Computing.* Retrieved from https://www.nsc.liu.se/lcsc2007/presentations/LCSC_2007-kelemen.pdf

National Digital Stewardship Alliance. (2013). NDSA Levels of Preservation. Retrieved from http://digitalpreservation.gov/ndsa/activities/levels.html

Rinehart, A. & Prud'homee, P. (2014). Overwhelmed to action: digital preservation challenges at the under-resourced institution. *OCLC Systems & Services, 30*(1), 28-42. doi:10.1108/OCLC-06-2013-0019

Rosenthal, D. S. H. Bit Preservation: A Solved Problem? *International Journal of Digital Curation, 5*(1), 134-148. doi: 10.2218/ijdc.v5i1.148

Wilsey, L., Skirvin, R., Chan, P. & Edwards, G. (2013). Capturing and Processing Born-Digital Files in the STOP AIDS Project Records: A Case Study. *Journal of Western Archives, 4*(1), 1-22.

Table of Figures

Table of Tables