Providing Access to Disk Image Content: A Preliminary Approach and Workflow

Walker Sampson
CU Boulder
University Libraries, 184 UCB
Boulder, CO 80309-0184
303-492-9161
walker.sampson@colorado.edu

Alexandra Chassanoff
UNC Chapel Hill
CB #3360, 100 Manning Hall
Chapel Hill, NC 27599-3360
919-962-8366
achass@email.unc.edu

ABSTRACT

In this poster, we describe a proposed workflow that can be used by collecting institutions acquiring disk images to support the provisioning of access to the born-digital content therein.

General Terms

Preservation strategies and workflows.

Keywords

Digital forensics; born-digital media; disk images.

1. INTRODUCTION

Born-digital materials are increasingly acquired by libraries, archives, and museums (LAMs). Though institutions have long been tasked with the preservation of collected materials, along with their continual access, born-digital data from removable media presents certain challenges [1]. One promising approach gaining traction among LAMs has been the adoption and use of open-source digital forensics software environments like BitCurator, for the capture and analysis of these born-digital materials [2].

However, there is currently limited support for institutions seeking to provide access to forensically captured born-digital content and associated metadata. The BitCurator Access project, which began in October 2014, seeks to address this gap by developing software to simplify access to content on raw and forensically packaged disk images.[1] In this poster, we propose a workflow that describes the capture, analysis, and final access to disk image content for collections held at the research archives at the University of Colorado Boulder.

2. BACKGROUND

The Archives at the University of Colorado Boulder has collected a wide range of floppy disk types; these reside in boxed folders or containers throughout its stacks. The Archives receives floppy disks as part of new accessions as well. While plans are in place for the implementation of both, the Archives has no software deployed which may function as a digital repository (e.g., DSpace, Fedora, Archivematica, Islandora), or collection management software deployed (e.g., ArchivesSpace, AtoM, PastPerfect).

The BitCurator Access project is currently developing BitCurator Access Web Tools (or bca-webtools) for web-based access to disk images.[2] Provision of access to both disk images and associated metadata through bca-webtools will help institutions capture and provide an access environment that reflects original order and relevant environmental context for collection materials. Additionally, the project proposes a fourth area of investigation related to access - the development of tools to aid in redacting sensitive data from disk images and other digital collections.

3. WORKFLOW

3.1 Goals and Context

The preliminary workflow described here addresses the immediate needs of the material, such as bit-level capture and triage, while remaining flexible enough to have the outputs integrate with a future digital repository and collection management software. We hope this approach allows the described methods a wider institutional applicability.

The workflow enables researchers to access a bit-level copy of a floppy disk found in an archival collection. Access is typically regarded as the last milestone of processing work, so the workflow strives for completeness to this point.

3.2 Overview of Proposed Workflow

The proposed workflow at the University of Colorado Boulder for processing born-digital materials will begin with obtaining the physical disk. The source media will be photographed and the archivist will begin the disk image acquisition process. Creation of the disk image can be performed through a number of devices, such as a USB-attached 3.5” disk drive in the case of the many IBM PC-formatted disks, or through floppy drive controllers such as the FC5025 for 5.25” disks and the KryoFlux controller in the case of either 3.5” or 5.25” disks.

Once the image has been created, a simple mount test will be run in the BitCurator environment. Floppy disk images might not mount for a variety of reasons including bad sectors, unknown file system types, or poor reads. Disks that are not mountable will be problematic for the next step, so these images will either be documented and set aside or resolved before further processing.

The BitCurator Reporting Tool will then be run to generate analytic reports on disk image content, including reporting on file formats and deleted files. The Reporting tool produces a DFMXL output through *fiwalk,* which is broadly analogous to a top-level inventory of disk image contents, and a PREMIS description. Other programs or processes can be carried out here as well, such as virus scans, and their outputs logged.

[1] http://access.bitcurator.net/index.php?title=Main_Page

[2] https://github.com/BitCurator/bca-webtools

The use of Simson Garfinkel's *bulk extractor* program, integrated with the BitCurator Reporting Tool through the *BEViewer* graphic front-end, reports on personally identifiable information and other sensitive content [3]. Information which a donor may have indicated should remain private can be discerned.

At this stage, the full context and content of the disk image is considered captured and described. The total output — disk image, logs of the disk imaging, a photograph of the media, and associated metadata and reports from the BitCurator Reporting Tool, will be placed into a single BagIt package and uploaded to a managed storage space with redundant copies.

The ability to control access to sensitive materials found on disk images is an explicit goal of the BitCurator Access project. The bca-webtools interface will use authentication at the local level to limit access to those materials flagged as containing potential PII in the previous step.

In this workflow, the aforementioned BagIt bag will become the formal archival information package (AIP) [3] in a designated repository at a future date. While implementation details are likely to change as the software develops, the workflow will place another copy of the disk image and attendant metadata in a location accessible to bca-webtools. We note here that the attendant metadata will likely be a subset, rather than a full copy, of the metadata and inventories available in the AIP. Even in the case of a disk image with no PII marked for redaction, unallocated user data extant after delete commands or overwrites may often prevent the full index present in the DFXML, or other such reports, to be available to the end user through bca-webtools. The precise relationship between the information contained in the suite of descriptive documents in the AIP, and the metadata used by bca-webtools and the end user, is subject to development.

The disk image interface can then provide access to the broader public, serving as a dissemination information package (DIP). Researchers will be able to browse and download the contents of the disk image through the software's web interface. This access point can be pointed to or indexed from a number of finding aid types, ranging from full EAD documents and library catalog entries to more custom online inventories.

4. CONCLUSION

In this poster, we describe how one institution – the University of Colorado Boulder – proposes to integrate digital forensics tools into their processing workflow to provide web-based access to disk image content. Although this poster describes a specific implementation, we anticipate that other institutions will likely follow similar steps in their workflow processing born-digital materials.

5. ACKNOWLEDGEMENTS

The BitCurator Access project is funded through the Andrew W. Mellon Foundation. The principle investigator is Christopher Lee and the co-principle investigator and technical lead is Kam Woods. Our software developer is Sunitha Misra.

6. REFERENCES

[1] Kirschenbaum, M.G., Ovenden, R., and Redwine, G. 2010. Digital forensics and born-digital content in cultural heritage collections. Council on Library and Information Resources. Washington DC. http://www.clir.org/pubs/reports/reports/pub149/pub149.pdf

[2] Lee, C.A., Woods, K., Kirschenbaum, M. and Chassanoff, A. 2013. *From bitstreams to heritage: Putting digital forensics into practice in collecting institutions.* White paper. University of North Carolina at Chapel Hill. http://www.bitcurator.net/wp-content/uploads/2013/11/bitstreams-to-heritage.pdf

[3] Garfinkel, S.L. Digital media triage with bulk data analysis and bulk extractor. February 2013. *Computer security*, 32(C):56–72.

[3] http://public.ccsds.org/publications/archive/650x0m2.pdf