A Demonstration of BitCurator Access Webtools and Disk Image Redaction Tools

Christopher A. Lee and Kam Woods
School of Information and Library Science
University of North Carolina
216 Lenoir Drive, CB #3360
1-(919)-966-3598
callee@ils.unc.edu; kamwoods@email.unc.edu

ABSTRACT
BitCurator Access is developing open-source software that supports the provision of access to disk images through three exploratory approaches: (1) building tools to support web-based services, (2) enabling the export of file systems and associated metadata, (3) and the use of emulation environments. This demonstration will highlight two BitCurator Access software products: BitCurator Access Webtools which supports browser-based search and navigation over data from disk images, and a set of scripts to redact sensitive data from disk images.

Categories and Subject Descriptors
H.3.7 [**Information Storage and Retrieval**]: Digital Libraries – *collection, dissemination, systems issues.*

General Terms
Provenance; Data Triage; Digital Forensics.

Keywords
Digital forensics; preservation; DFXML; metadata; privacy; collections; web access; redaction

1. BITCURATOR ACCESS PROJECT
The BitCurator Access project began on October 1, 2014 and will end on September 30, 2016. Funded through a grant from the Andrew W. Mellon Foundation, BitCurator Access is developing open-source software that supports the provision of access to disk images through three exploratory approaches: (1) building tools to support web-based services, (2) enabling the export of file systems and associated metadata, (3) and the use of emulation environments. Also closely associated with these access goals is redaction. BitCurator Access is developing tools to redact files, file system metadata, and targeted bitstreams within disks or directories.

BitCurator Access focuses on approaches that simplify access to raw and forensically-packaged disk images; allowing collecting institutions to provide access environments that reflect as closely as possible the original order and environmental context of these materials. The use of forensic technologies allows for detailed metadata to be generated to reflect the provenance of the materials, the exact nature of the file-level items they contain, and the metadata associated with both file-level items and data not observed within the file system (but still accessible within the original materials). We are freely disseminating the BitCurator Access software products under an open source (GPL, Version 3) license. All existing software upon which the products are built is also either open-source or public domain.

This demonstration will highlight two BitCurator Access software products: BitCurator Access Webtools which supports browser-based search and navigation over data from disk images, and a set of scripts to redact sensitive data from disk images. We have previously reported on support for workflows that employ BCA Webtools and Emulation-as-a-Service (EaaS) [3].

2. BITCURATOR ACCESS WEBTOOLS

The BitCurator Access project has developed BCA Webtools, which is a suite of software (based on an earlier prototype called DIMAC [2]) that allows users to browse a wide range of file systems contained within disk images using a web browser. It is intended to support access requirements in libraries, archives, and museums preserving born-digital materials extracted from source media as raw or forensically-packaged disk images.

BCA Webtools uses open source libraries and toolkits including The Sleuth Kit, PyTSK, and the Flask web microservices framework. It uses PyLucene along with format-specific text-extraction tools to index the contents of files contained in disk images, allowing users to search for relevant content without individually inspecting files. BCA Webtools is distributed with a simple build script that deploys it as a Vagrant virtual machine running the web service.

The application can parse raw and E01-packaged images containing FAT16, FAT32, NTFS, HFS+, and EXT 2/3/4 file systems, and allows users to navigate the file system contents, download individual files, and search the contents within a simple web interface.

3. REDACTION TOOLS
Digital media acquisitions in libraries, archives and museums often contain data that may be classified as private, sensitive, or individually identifying, and the complexity and volume of information being collected demands automation to ensure that risks of inadvertent disclosure are minimized.

Currently, there are relatively few open source redaction tools capable of addressing these needs. BitCurator Access is target specific areas of software development, including:

- Redacting specific bitstreams from raw disk images

- Creating redacted copies of forensically-packaged disk images
- Building redaction overlays that can applied to disk images in an access context, masking out specific files and directories
- Redacting metadata from commonly used file formats, including Office and PDF files.

This demonstration will include modifications to and adaptations of existing Digital Forensics XML tools [1] that provide support for the above activities. Specifically, we will demonstrate a Python tool for redacting sequences of data from disk images matching one or more pattern(s) provided as arguments on the command line or in a configuration file.

The demonstrated redaction tool that is neither file system nor file format sensitive by default, although it may operate using the output of tools that output file system statistics including byte runs associated with individual files and directories identified within recognized file systems. The tool will also perform redaction operations on relevant byte sequences identified in raw data streams, whether or not they are presented in the form of disk images.

4. ACKNOWLEDGMENTS

Development of the BitCurator environment and BitCurator Access Webtools have been supported by the Andrew W. Mellon Foundation. The BitCurator Access team is composed of Alex Chassanoff, Christopher Lee, Sunitha Misra, and Kam Woods. Greg Jansen contributed significantly to the development and documentation of the redaction tools featured in this demonstration.

5. REFERENCES

[1] Garfinkel, S. Digital Forensics XML and the DFXML Toolset. *Digital Investigation* 8 (2012), 161-174.

[2] Misra, S., Lee, C. A., and Woods, K. 2014. A Web Service for File-Level Access to Disk Images. *Code4Lib Journal* 25 (2014).

[3] Woods, K., Lee, C. A., Stobbe, O., Liebetraut, T., and Rechert, K. 2015. Functional Access to Forensic Disk Images in a Web Service. In *Proceedings of iPRES 2015*. University of North Carolina, Chapel Hill, NC, 191-195.