

Acquiring and Providing Access to Born-Digital Materials with the BitCurator Environment and BitCurator Access Webtools

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

ABSTRACT

This tutorial will prepare participants to use the open-source BitCurator environment and BitCurator Access Webtools to acquire, process and provide access to born-digital materials. There will be a brief lecture and discussion that focuses on the motivation for using the tools and several foundational technical concepts. The remainder of the tutorial will be devoted to demonstration and hands-on exercises that feature specific tools and methods. Participants will learn how to mount media as read-only, create disk images, mount forensically packaged disk images, export individual files or entire directories from disk images, use specialized scripts to perform batch activities, generate and interpret Digital Forensics XML (DFXML), generate a variety of standard and customized reports (including PREMIS records), identify various forms of sensitive data within collections, and provide browser-based search and navigation of files and folders.

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

Forensics; preservation; DFXML; metadata; privacy; collections; acquisition; web access

1. BITCURATOR PROJECT

The BitCurator Project, a collaborative effort led by the School of Information and Library Science at the University of North Carolina at Chapel Hill and Maryland Institute for Technology in the Humanities at the University of Maryland, is addressing two

iPres 2016 conference proceedings will be made available under a Creative Commons license.

With the exception of any logos, emblems, trademarks or other nominated third-party images/text, this work is available for reuse under a Creative Commons Attribution 3.0 unported license. Authorship of this work must be attributed. View a [copy of this licence](#).

fundamental needs and opportunities for collecting institutions: (1) integrating digital forensics tools and methods into the workflows and collection management environments of libraries, archives and museums and (2) supporting properly mediated public access to forensically acquired data [4].

2. BITCURATOR ENVIRONMENT

We are developing and disseminating a suite of open source tools. These tools are being developed and tested in a Linux environment; the software on which they depend can readily be compiled for Windows environments (and in most cases are currently distributed as both source code and Windows binaries). We intend the majority of the development for BitCurator to support cross-platform use of the software. We are freely disseminating the software under an open source (GPL, Version 3) license. BitCurator provides users with two primary paths to integrate digital forensics tools and techniques into archival and library workflows.

First, the BitCurator software can be run as a ready-to-run Linux environment that can be used either as a virtual machine (VM) or installed as a host operating system. This environment is customized to provide users with graphic user interface (GUI)-based scripts that provide simplified access to common functions associated with handling media, including facilities to prevent inadvertent write-enabled mounting (software write-blocking).

Second, the BitCurator software can be run as a set of individual software tools, packages, support scripts, and documentation to reproduce full or partial functionality of the ready-to-run BitCurator environment. These include a software metapackage (.deb) file that replicates the software dependency tree on which software sources built for BitCurator rely; a set of software sources and supporting environmental scripts developed by the BitCurator team and made publicly available at via our GitHub repository (links at <http://wiki.bitcurator.net>); and all other third-party open source digital forensics software included in the BitCurator environment.

3. BITCURATOR ACCESS WEBTOOLS

The BitCurator Access project has developed BCA Webtools, which is a suite of software 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.

4. TUTORIAL FORMAT

There will be a brief lecture and discussion that focuses on the motivation for using the tools and several foundational technical concepts. The remainder of the tutorial will be devoted to demonstration and hands-on exercises that demonstrate specific tools and methods.

5. INTENDED AUDIENCE

This tutorial should be of interest to information professionals who are responsible for acquiring or transferring collections of digital materials, particularly those that are received on removable media. Another intended audience is individuals involved in digital preservation research, development and IT management, who will learn how data generated within the BitCurator environment can complement and potentially be integrated with data generated by other tools and systems.

6. EXPECTED LEARNING OUTCOMES

This tutorial will prepare participants to use the open-source BitCurator environment to acquire and process born-digital data. Tools that BitCurator is incorporating include Guymager, a program for capturing disk images; bulk extractor, for extracting features of interest from disk images (including private and individually identifying information); fiwalk, for generating Digital Forensics XML (DFXML) output describing filesystem hierarchies contained on disk images; The Sleuth Kit (TSK), for viewing, identifying and extraction information from disk images; Nautilus scripts to automate the actions of command-line forensics utilities through the Ubuntu desktop browser; and sdhash, a fuzzing hashing application that can find partial matches between similar files. For further information about several of these tools, see [1,2,3,5].

Upon completion of this tutorial, participants should understand several of the major motivations and uses cases for applying the BitCurator environment. They will also know how to perform the following tasks:

- mount media as read-only
- create disk images, mount forensically packaged disk images

- export individual files or entire directories from disk images
- use specialized scripts to perform batch activities
- generate and interpret Digital Forensics XML (DFXML) generate a variety of standard and customized reports (including PREMIS records)
- identify various forms of sensitive data within collections
- provide browser-based search and navigation of files and folders.

Participants will also become aware of the resources that are available for learning more about the software and engage with other users after completion of the tutorial.

7. INSTRUCTOR BIOGRAPHY

Christopher (Cal) Lee is Associate Professor at the School of Information and Library Science at the University of North Carolina, Chapel Hill. He teaches graduate and continuing education courses in archival administration, records management, digital curation, and information technology for managing digital collections. His research focuses on curation of digital collections and stewardship of personal digital archives. Cal is PI for the BitCurator project and editor of *I, Digital: Personal Collections in the Digital Era*.

8. ACKNOWLEDGMENTS

Development of the BitCurator environment and BitCurator Access Webtools have been supported by the Andrew W. Mellon Foundation.

9. REFERENCES

- [1] Cohen, M., Garfinkel, S., and Schatz, B. 2009. Extending the Advanced Forensic Format to Accommodate Multiple Data Sources, Logical Evidence, Arbitrary Information and Forensic Workflow. *Digital Investigation* 6 (2009), S57-S68.
- [2] Garfinkel, S. Digital Forensics XML and the DFXML Toolset. *Digital Investigation* 8 (2012), 161-174.
- [3] Garfinkel, S.L. Providing Cryptographic Security and Evidentiary Chain-of-Custody with the Advanced Forensic Format, Library, and Tools. *International Journal of Digital Crime and Forensics* 1, 1 (2009), 1-28;
- [4] Lee, C.A., Kirschenbaum, M.G., Chassanoff, A., Olsen, P., and Woods, K. BitCurator: Tools and Techniques for Digital Forensics in Collecting Institutions. *D-Lib Magazine* 18, 5/6 (May/June 2012).
- [5] Roussev, V. An Evaluation of Forensic Similarity Hashes. *Digital Investigation* 8 (2011), S34-S41.