Acquiring and Processing Born-Digital Data Using the BitCurator Environment

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 to acquire and process born-digital data. 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. 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 Nautilus scripts to perform batch activities, generate and interpret Digital Forensics XML (DFXML), generate a variety of standard and customized reports (including PREMIS recods), and identify various forms of sensitive data within collections.

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

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 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].

iPres 2014 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 re-use under a Creative Commons Attribution 3.0 unported license. Authorship of this work must be attributed. View a copy of this licence.

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. TUTORIAL FORMAT
This is being proposed as a full-day (6-hour) 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.

4. 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.

5. 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 Nautilus 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.

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.

6. 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*.

7. ACKNOWLEDGMENTS

BitCurator development has been supported by the Andrew W. Mellon Foundation. Members of the BitCurator team are Alexandra Chassanoff, Matthew Kirschenbaum, Christopher (Cal) Lee, Sunitha Misra, Porter Olsen, and Kam Woods. Members of two advisory boards have made valuable contributions: the Development Advisory Group (DAG) and Professional Experts Panel (PEP).

8. 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.