Integrating Digital Forensics into Born-Digital Workflows: The BitCurator Project

Martin Gengenbach[1], Alexandra Chassanoff[1], Porter Olsen[2]

1 School of Information and Library Science, University of North Carolina Chapel Hill 2 Maryland Institute for Technology in the Humanities, University of Maryland

INTRODUCTION

Digital forensics (DF) is defined as an applied field "concerned with discovering, authenticating, and analyzing data in digital formats to the standard of admissibility in a legal setting" [1]. There is a rapidly developing body of information on the ways in which DF tools, practices, and technologies may be used on born-digital materials [2, 3, 4, 5], as well as evolving work on best practices for the effective management of born-digital collections [6, 7]. However, there have been relatively few case studies on the implementation of such tools in working institutional settings. We will attempt to address that gap by presenting the preliminary results of interviews conducted with collecting institutions in the context of the BitCurator Project, a suite of DF tools being developed for the library/archives community.

BitCurator is a Mellon-funded initiative led by the School of Information and Library Science (SILS) at the University of North Carolina at Chapel Hill and Maryland Institute for Technology in the Humanities (MITH) at the University of Maryland.

The project has three major goals:

- to develop and test a system for using forensics methods in collecting institutions;
- to define a workflow that implements these methods;
- to support properly mediated public access to forensically acquired data.

METHODOLOGY

In the first phase of the project, BitCurator assembled two advisory groups to discuss software requirements, review design assumptions, draft institutional workflows, and help scope project goals. The **Development Advisory Group (DAG)** is comprised of technologists with specialties in digital preservation, digital archives, and digital forensics. The **Professional Experts Panel (PEP)** is made up of professional archivists and librarians from institutions that are acquiring and preserving born-digital materials. Following those initial discussions, the project team drafted an overview of BitCurator functionalities that would complement, support, and enhance existing digital curation workflows.

We conducted in-depth, semi-structured interviews with 8 members of the DAG and PEP advisory boards, plus one archivist with experience implementing digital forensic workflows. We drafted institutional workflows based on the product of those interviews, which also provided participants the opportunity to reflect upon the successes and challenges of their current workflows for managing born-digital content.

CAPTURING INSTITUTIONAL WORKFLOWS

Participants interviewed noted similar issues in current born-digital workflows. Relevant findings include the following:

- **Forensic disk images (bit-level copies) are created and used in born-digital workflows**. Participants cited the importance of stabilizing the image and maintaining an "archival snap shot" of the content to be ingested into preservation storage.
- **Workflows need to be simplified.** Initial accessioning and capture activities, including imaging and extraction of basic metadata, can quickly overwhelm an archivist.
- **Discovery and redaction of personal identification information (PII) should be automated.** All participants described current PII processes as arduous and unrealistic for collections of any significant size and complexity.
- **Tools must be flexible, modular, and scalable.** Participants acknowledged that their workflow process was subject to constant re-evaluation in the context of emerging technologies and best practices, and shifting funding and institutional priorities. They articulated a need for tools to be able to accommodate varied institutional and technological contexts, providing standardized outputs in widely accepted formats to integrate with current institutional technologies.

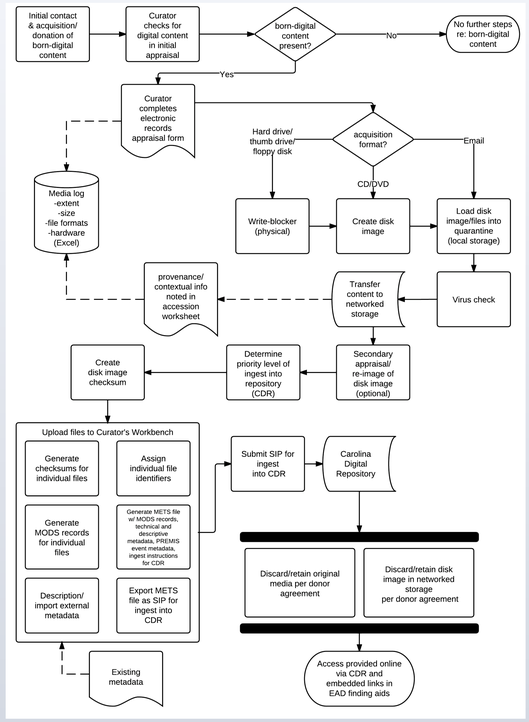

Figure 1: Born-digital workflow, University of North Carolina at Chapel Hill.

SUPPORTING CURATION TASKS THROUGH BITCURATOR FUNCTIONALITIES

By bringing together a suite of existing open source DF applications, BitCurator is able to leverage work already being done in the community, as well as create an environment specifically focused on the needs of digital archivists. The BitCurator environment supports functionalities within three major categories:

- ***Acquisition*** - BitCurator assists the digital archivist with calculating check sums, initial data triage, and file search on the original disk.
- ***Staging and Pre-Ingest*** - BitCurator uses Bulk Extractor to automate the identification of PII and then outputs that data in a human readable format, allowing the digital archivist to decide what data to exclude before the disk is imaged.
- ***Ingest and Archival Storage*** - BitCurator facilitates the storage of archival metadata that remains part of the forensic disk image.

In addition, BitCurator will produce forensics-level reports in the Digital Forensics XML (DFXML) schema, which can then be cross-walked to METS or other metadata encoding standards.

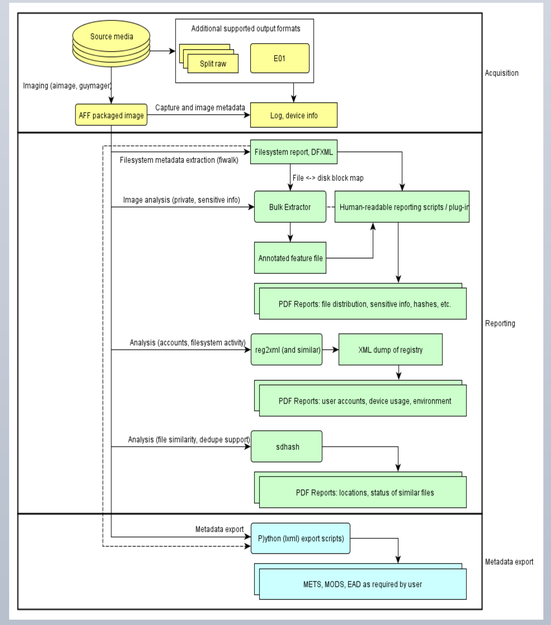

Figure 2: Overview of BitCurator-supported functionalities

FUTURE WORK

BitCurator is intended to be a two-phase project, with the first phase occurring in years one and two. In this poster, we have reported on results obtained in the first year of the project which include: (1) detailed workflows documenting the handling of born-digital content in several collecting institutions; (2) specifications on how BitCurator can support the implementation of digital forensics tools and methods in curatorial workflows. Using a community-driven model, future work will explore how to capitalize on the granularity and richness of the information captured by DF tools with appropriate arrangement and description of, and access to, those materials.

REFERENCES

[1] Kirschenbaum, M.G., Ovenden, R., & Redwine, G. (2010). *Digital forensics and born-digital content in cultural heritage collections*. Council on Library and Information Resources.

[2] Duranti, L. (2009). From digital diplomatics to digital records forensics. *Archivaria* 68, 39-66.

[3] Garfinkel, S., & Cox, D. (2009, February 9-11). *Finding and archiving the internet footprint*. Paper presented at the First Digital Lives Research Conference: Personal Digital Archives for the 21st Century, London, UK.

[4] John, J. L. (2008). *Adapting existing technologies for digitally archiving personal lives: Digital forensics, ancestral computing, and evolutionary perspectives and tools*. In Fifth International Conference on Preservation of Digital Objects, London, UK.

[5] Woods, K, Lee, C.A., & Garfinkel, S. (2011). Extending digital repository architectures to support disk image preservation and access. In *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries* (pp.57-66). New York, NY: Association for Computing Machinery.

[6] AIMS Working Group (2012). *AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship*.

[7] Thibodeau, K. (2002) Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years. In *The State of Digital Preservation: An International Perspective, 4-31: Council on Library and Information Resources*.

ACKNOWLEDGEMENTS

The BitCurator project is funded through the Andrew W. Mellon Foundation. The principle investigator is Christopher Lee and the co-principle investigator is Matthew Kirschenbaum. BitCurator's technical lead is Kam Woods and our software developer is Sunitha Misra. We would like to thank all the members of the PEP and DAG advisory groups for their contributions to this project.

RESEARCH POSTER PRESENTATION DESIGN © 2012
www.PosterPresentations.com