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

Martin Gengenbach  
School of Information & Library Science  
University of North Carolina at Chapel Hill  
martin.gengenbach@gmail.com

Alexandra Chassanoff  
School of Information & Library Science  
University of North Carolina at Chapel Hill  
achass@email.unc.edu

Porter Olsen  
Maryland Institute of Technology  
University of Maryland  
polsen@umd.edu

ABSTRACT
There is a growing body of work investigating the needs and desires of collecting institutions as they adapt to the acquisition of born-digital materials. The incorporation of digital forensics tools and techniques into digital curation workflows offers great promise for addressing the complexities bound up in ingesting and preserving digital objects at multiple levels of representation. This poster presents preliminary results from ongoing research conducted as part of the BitCurator project, a two-year grant funded initiative to build, test, and analyze systems and software for incorporating digital forensics methods into collecting institutions’ workflows. The project arose out of a perceived need in the library, archives, and museum (LAM) communities for better documentation, interfaces, and functionality in processing born-digital archival materials.

Keywords  
Digital forensics, digital archives, digital curation.

INTRODUCTION
The acquisition of archival born-digital materials by collecting institutions has grown exponentially in recent years. According to a 2010 survey on ARL Research Libraries Special Collections and Archives, 79% of respondents surveyed had collected born-digital materials in one or more formats (Dooley & Luce, 2010, p. 59).

Despite the growth in acquisition of digital formats, the management of born-digital materials present numerous challenges for cultural heritage institutions / libraries, archives, and museums (LAM). Archivists, curators, and librarians confront a host of technical issues when handling digital materials, including recovery of legacy media formats, file system discrepancies, and hardware risks (Kirschenbaum, Ovenden, & Redwine, 2010, pp. 14-21). Institutions must be able to effectively integrate tools and technologies for processing born-digital materials into their workflow.

The application and use of digital forensics (DF) techniques to the processing and curation of digital materials is one approach being explored by cultural heritage institutions. 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, as well as evolving work on best practices for the management of born-digital collections. However, there have been relatively few specific 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: (1) to develop and test a system for using forensics methods in collecting institutions; (2) to define a workflow that implements these methods; (3) to support properly mediated public access to forensically acquired data.

LITERATURE REVIEW
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” (Kirschenbaum, Ovenden, & Redwine, 2010, p.1). The application of DF in digital preservation and
digital curation has developed in tandem with the recognition that digital objects are complex entities with multiple inheritances, depending on their interpretation as physical, logical, or conceptual objects (Thibodeau, 2002); that digital media, though fragile, can retain significant amounts of stored data (Ross and Gow, 1999); and that archivists, as with paper records, must be able to verify the authenticity of the content they manage in digital form (Duranti, 1995, p.8).

Previous work has focused on identifying potential uses for digital forensic tools in cultural heritage settings (Garfinkel & Cox, 2009; John, 2009) and grounding digital forensics in contemporary archival theory and best practices (Duranti, 2009; Duranti & Endicott-Popovsky, 2010; Woods, Lee, & Garfinkel, 2011). Additionally, several grant-funded projects have focused on or incorporated digital forensics into their research objectives: the British Library’s Digital Lives Research Project, for example, identifies digital forensic capture of disk information as a way of ensuring the authenticity of data (John et al., 2009, p.126); a conference funded by the Council on Library and Information Resources produced Digital Forensics and Born-Digital Content in Cultural Heritage Collections, a report that directly considered the challenges that born-digital content poses to collecting institutions (Kirschchenbaum, 2010); and recently the AIMS project has developed an inter-institutional model for stewardship of born digital collections, using and standardizing institutional models of digital forensic capture, stabilization, and provision of access (AIMS Project Group, 2012).

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 put together a draft outline of a BitCurator-supported institutional workflow. The intent of this document was to demonstrate the functionalities offered by BitCurator to complement, support, and enhance existing digital curation workflows.

In related research, a Master's Student at the School of Information and Library Science at the University of North Carolina at Chapel Hill conducted in-depth interviews with 8 members from the DAG and PEP advisory boards, plus one archivist with experience implementing digital forensic workflows. He conducted semi-structured interviews and drafted institutional workflows based on the product of those interviews, following up with participants to clarify and refine the resulting documents. The interviews also provided participants the opportunity to reflect upon the successes and challenges of their current workflows, changes they would like to make, and what they had learned through the creation and evolution of their current processes.

PRELIMINARY FINDINGS
Capturing institutional workflows
Participants who were interviewed from the PEP and DAG groups were all operating in their home institutions with a bit-copy (disk image) at the heart of their accessioning and preservation efforts. In cases where a raw disk image is currently being used, the AFF disk image is a tempting path because of storage considerations and metadata packaging. However, participants also noted that the initial accessioning and capture activities, including imaging and extraction of basic metadata, can quickly overwhelm an archivist. Many interviewees mentioned the need to streamline and simplify their workflow. Participants expressed interest in automating the discovery and redaction of personal identifiable information (PII). This, too, can be a time-consuming, arduous activity, and in most cases unrealistic for a collection of any significant size and complexity. Automating as many steps as possible in the process was mentioned by participants as an important feature for any tool, as well as making it accessible for paraprofessional and non-technical staff, students, or volunteers.

All participants acknowledged that their workflow process was subject to continuous change and re-evaluation in the context of available technologies, emerging best practices, and shifting funding and institutional priorities. Therefore, participants recommended that tools developed to accession digital content must be similarly flexible, modular, and scalable. Similarly, there is likely not a single unified workflow for every institution. Born-digital content workflows exist in different institutional and technological contexts, where many different collections may be acquiring born-digital materials. Thus participants articulated a need for tools to be able to accommodate such inconsistencies while providing standardized outputs in widely accepted formats to integrate with current institutional technologies.

BitCurator-supported functionalities
BitCurator can best be thought of as an environment rather than a single, monolithic application. By bringing together a suite of existing open source digital forensics (DF) applications, BitCurator is able to leverage work already being done in the DF community and also create an environment specifically focused on the needs of digital archivists.

The functionalities of the BitCurator environment fall within three major categories: 1) Acquisition, 2) Staging and Pre-Ingest, and 3) Ingest and Archival Storage.
Figure 1: UNC-Chapel Hill born-digital workflow
During Acquisition BitCurator assists the digital archivist with tasks such as calculating checksums, initial data triage through file type identification, and the ability to quickly search for files on the original disk (see Media log in figure 1 above).

In Staging and Pre-Ingest BitCurator works with 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.

During Ingest and Archival Storage BitCurator will facilitate the storage of archival metadata that remains part of the forensics disk image. In addition, BitCurator will produce forensics level reports via DF tools Fiwalk, Bulk Extractor and TSK. These reports will be produced in Digital Forensics XML (DFXML) form, which can then be cross-walked to METS or other metadata encoding standards.

IMPLICATIONS AND 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 workflow. 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/access to those materials.

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

REFERENCES


