PRINTER-FRIENDLY FORMAT Return to Article

D-Lib Magazine

May/June 2012
Volume 18, Number 5/6

BitCurator: Tools and Techniques for Digital Forensics in Collecting Institutions

Christopher A. Lee, University of North Carolina, Chapel Hill
Matthew Kirschenbaum, University of Maryland
Alexandra Chassanoff, University of North Carolina, Chapel Hill
Porter Olsen, University of Maryland
Kam Woods, University of North Carolina, Chapel Hill

doi:10.1045/may2012-lee

Abstract

This paper introduces the BitCurator Project, which aims to incorporate digital forensics tools and methods into collecting institutions' workflows. BitCurator is a collaborative effort 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 arose from a perceived need in the library/archives community to develop digital forensics tools with interfaces, documentation, and functionality that can support the workflows of collecting institutions. This paper describes current efforts, ongoing work, and implications for future development of forensic-based, analytic software for born-digital materials.

1. Introduction

The acquisition of digital materials by collecting institutions — libraries, archives and museums (LAMs) — has resulted in the need to incorporate new tools and methods into curatorial practices. LAMs are increasingly called upon to move born-digital materials from removable media into more sustainable preservation environments. This can involve media that are already in their holdings (e.g., disks stored in boxes alongside paper materials), as well as materials being acquired for the first time from individual donors or other producers.

Digital collections can be encountered at multiple levels of representation, which brings unprecedented opportunities for description, interpretation and use. There is a substantial body of information within the data structures of computer systems that can often be discovered or recovered. Along with new opportunities, institutions also face a variety of technical difficulties as they process digital data, including: filesystem discrepancies between storage environments, media format obsolescence, operating system incompatibilities, and hardware risks (Kirschenbaum, Ovenden, & Redwine, 2010, pp. 14-21).

Integrating digital forensics approaches with curation workflows can help LAM professionals to ensure the authenticity, integrity, and provenance of digital materials. Digital forensics focuses on the discovery, recovery, and validation of information from computer systems — information that is often not immediately visible to common users. More than a decade ago, a report by Seamus Ross and Ann Gow (1999) discussed the potential relevance of advances in data recovery and digital forensics to collecting institutions. More recently, there has been an active stream of literature related to the use of forensic tools and methods for acquiring and managing digital collections (Duranti, 2009; Duranti & Endicot-Popovsky, 2010; Elford *et al.*, 2008; Garfinkel & Cox, 2009; John, 2008; Underwood & Laib, 2007; Underwood *et al.*, 2009; Woods & Brown, 2009; Woods, Lee & Garfinkel, 2011; Xie, 2011). A project called "Computer Forensics

and Born-Digital Content in Cultural Heritage Collections" hosted a symposium and generated a report in 2010. Both the symposium and the report provided significant contributions to this discussion. The Born Digital Collections: An Inter-Institutional Model for Stewardship (AIMS) project developed a framework for the stewardship of born-digital materials that includes the incorporation of digital forensics methods (AIMS Work Group, 2012). The Library of Congress named the growing visibility of digital forensics as one of the Top Ten digital preservation developments of 2010.

The BitCurator Project, a collaborative effort 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, builds on previous work by addressing two fundamental needs and opportunities for collecting institutions: (1) integrating digital forensics tools and methods into the workflows and collection management environments of LAMs and (2) supporting properly mediated public access to forensically acquired data. The project team has developed a set of requirements documents, which we are iteratively revising based on input from our expert advisors. We are currently developing prototype software that will be subject to various forms of testing and review.

2. Summary of DAG and PEP Panel Meetings

The BitCurator project has enlisted the support of two advisory groups: the Professional Experts Panel (PEP) and the Development Advisory Group (DAG). The core project team met with the PEP in December of 2011 and the DAG in January of 2012 to discuss the design assumptions and goals of the project. An overview of each advisory group's role and contribution is outlined below.

Professional Experts Panel

The PEP is composed of professionals working at collecting institutions; these include local municipalities, national government archives, academic institutions, and libraries. Each panel member has practical experience working in digital curation, and their respective institutions are all exploring or implementing digital forensics tools and methods.

In December of 2011, PEP members gathered with the BitCurator team for a two-day meeting at MITH to discuss initial drafts of BitCurator design documents, identify needs not currently met by existing open source digital forensics tools, and discuss how BitCurator might complement, support and enhance existing digital curation workflows.

Critical PEP recommendations for BitCurator software included the need for the following:

- Clear, user-friendly documentation
- Modular, cross-platform software tools
- Software and guidance for collecting born-digital materials remotely at a donor's facility or residence
- An easily navigable graphical user interface, an application programming interface (API) for integration with existing software platforms, and command line tools that support batch processing
- Data triage functions to automate repetitive or technically challenging tasks during both appraisal and reprocessing after ingest (e.g., high-quality file type identification, flagging and redaction of private and sensitive data, and general reports on drive contents).

PEP participants outlined and documented their existing collection acquisition and archival processing workflows. As expected, there were significant variations in the workflow practices of the respective organizations. During this session we attempted to capture the representative digital curation workflows (and map their similarities and differences) by having PEP members describe and sketch both their "official" workflow and those steps, issues, quirks, and undocumented detours that result from everyday practice. These workflow representations provide a concrete guide for our ongoing work. As part of the BitCurator development process, we are refining a master map that identifies parallel areas in the workflow groups, identifies process limitations (or gaps), and links to novel functionality provided by the tools.

Development Advisory Group

The DAG is composed of academic, government and industry professionals who have significant experience with development or management of development groups. The majority of these members work in, or closely with, collecting institutions. DAG members are assisting in BitCurator's design process and scoping, along with mapping in-scope items to specific needs. Ongoing communication with the DAG also will ensure that BitCurator efforts complement software being developed by others.

The first BitCurator DAG meeting coincided with an event on January 6, 2012 in Chapel Hill called CurateGear: Enabling the Curation of Digital Collections, where DAG members gave presentations on projects and activities within their own institutions. The DAG subsequently met for a full day to review BitCurator design documentation and objectives and to discuss development strategies and potential challenges. Review of documentation (including design documents and feedback from the PEP) comprised the bulk of the day's work, with DAG members addressing a wide range of issues, including:

- BitCurator's role in the broader ecology of digital archives tools
- Project scope, objectives and planned deliverables
- Defining intended user groups for BitCurator software
- The need for both GUI and command line interfaces, facilitating interactive and batch processing
- The need to identify and assess private and sensitive data during multiple stages of the curation process
- Education and documentation requirements
- Opportunities for collaboration with and among members of the DAG
- Outreach and long-term support and for project deliverables.

The meeting concluded with a discussion of the possible research projects and publications that might emerge in conjunction with the development of BitCurator.

3. Current and Ongoing Development Work

Several iterations of design documentation and scoping work have provided a baseline for software and standards that will provide a foundation for BitCurator development. Beyond the requirements outlined during the PEP meeting, our criteria include the following:

- The forensic technologies on which BitCurator relies should be open source, extensible, and mature.
- Development will focus on extensions, plugins, and wrappers for proven software rather than from-scratch development.
- BitCurator will adhere to a common digital forensic metadata standard and provide crosswalks to relevant library and archives metadata schemas.

A growing number of open source applications provide support for the Advanced Forensic Format (and associated libraries and utilities to read and write this format), and Digital Forensics XML (DFXML), an initiative to enable the production of interoperable metadata by digital forensics tools. We are adopting AFF and DFXML as core technologies for BitCurator. Current software development is focused on the construction of simplified reporting mechanism for outputs from the disk image parsing tool called *fiwalk* and the private and sensitive feature extraction mechanisms in *bulk extractor* (tools developed by Simson Garfinkel). Building upon these tools provides access to high-performance data stream analytics, an existing plugin mechanism, and API hooks in both C and Python.

The BitCurator mechanisms currently in development will enable the following:

- Simple overviews of all personally identifying information (PII) found by *bulk extractor*, at a level of granularity that can be controlled by the user, and output in both human-readable and machine-readable formats
- Batch extraction of all files from a disk image corresponding to a particular feature set (PII or criteria described by the user), to support assessment, archival description and redaction.

These tools are currently being developed and tested in a Linux environment (Ubuntu 11.10 at the time of writing), although the software on which they depend (the AFFLIB suite of libraries and tools, bulk extractor, fiwalk, and The Sleuth Kit) 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.

4. Implications and Future work

We have designed BitCurator as a two-phase project, with the first phase occurring in years one and two. The current grant from the Andrew W. Mellon Foundation provides funding for the first phase. If it is funded as the result of a follow-on proposal for phase two, the third year of the project will include deployment of the initial tool sets among both (1) the institutions represented by the Professional Experts Panel and (2) institutions (to be determined in year two) that will be testing a "BitCurator in a Box," a package of materials, including a write blocker, software environment installed on a bootable USB drive, and associated documentation to generate disk images of holdings and then transfer the disk images (if desired) to a BitCurator demonstration testbed environment for further processing. Using a community-driven model already successfully tested in MITH's recent TILE (Text-Image Linking Environment) project, we would work with the partners in phase two to iterate and refine the software, ensuring robust and usable documentation, and (finally) we would aggressively pursue community uptake through the efforts of a dedicated community lead who would actively promote and support the tools.

There are numerous elements of the BitCurator project that are designed to build capacity and ensure the sustainability of digital acquisition education activities. BitCurator is a joint effort between faculty and personnel from an Information School and a digital humanities center with expertise in software development. The BitCurator software will be distributed under an open source license. Diverse constituencies will thus be able to extend the tools. This alone, however, is not enough to ensure sustainability. We also aim to develop robust end-user documentation and provide support for community uptake as the tools are distributed and integrated into working collections and repositories. Finally, we intend to explore forensic archival practice in the context of emerging theoretical paradigms from media and cultural studies, notably media archeology (Huhtamo & Parikka, 2011).

5. References

[1] AIMS Work Group. (2012). *AIMS born-digital collections: An inter-institutional model for stewardship*. Retrieved from: http://www2.lib.virginia.edu/aims/whitepaper/AIMS_final.pdf.

[2] Duranti, L. (2009). From digital diplomatics to digital records forensics. *Archivaria* 68, 39-66. Retrieved from: http://journals.sfu.ca/archivar/index.php/archivaria/article/viewArticle/13229.

[3] Duranti, L., & Endicott-Popovsky, B. (2010). Digital records forensics: A new science and academic program for forensic readiness. *Journal of Digital Forensics, Security and Law*, 5(2). Retrieved from: http://www.jdfsl.org/subscriptions/JDFSL-V5N2-Duranti.pdf.

[4] Elford, D., Del Pozo, N., Mihajlovic, S., Pearson, D., Clifton, G., & Webb, C. (2008, August 10-14). *Media matters: Developing processes for preserving digital objects on physical carriers at the National Library of Australia*. Paper presented at the 74th IFLA General Conference and Council, Québec, Canada. Retrieved from: http://archive.ifla.org/IV/ifla74/papers/084-Webb-en.pdf.

[5] Garfinkel, S. L. (2010, August). Digital forensics research: The next 10 years. *Digital Investigation*, 7. Retrieved from: http://dx.doi.org/10.1016/j.diin.2010.05.009.

[6] 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. Retrieved from: http://simson.net/clips/academic/2009.BL.InternetFootprint.pdf.

[7] Huhtamo, E. & Parikka, J. (Eds). (2011). *Media Archeology: Approaches, Applications, and Implications*. Berkeley, California: University of California Press.

[8] John, J. L. (2008, September 29-30). *Adapting existing technologies for digitally archiving personal lives: Digital forensics, ancestral computing, and evolutionary perspectives and tools*. Paper presented at the iPRES 2008: The Fifth

International Conference on Preservation of Digital Objects, London, UK. Retrieved from: http://www.bl.uk/ipres2008/presentations_day1/09_John.pdf.

[9] Kirschenbaum, M.G., Ovenden, R., & Redwine, G. (2010). *Digital forensics and born-digital content in cultural heritage collections*. Council on Library and Information Resources. Retrieved from: http://www.clir.org/pubs/reports/pub149/reports/pub149/pub149.pdf.

[10] Ross, S., & Gow, A. (1999). *Digital archaeology: Rescuing neglected and damaged data resources, a JISC/NPO study within the electronic libraries (eLib) programme on the preservation of electronic materials*. London: Library Information Technology Center. Retrieved from: http://eprints.erpanet.org/47/.

[11] Underwood, W. E., & Laib, S.L. (2007, April 18-20). *PERPOS: An electronic records repository and archival processing system*. Paper presented at the International Symposium on Digital Curation (DigCCurr 2007), Chapel Hill, NC. Retrieved from: http://ils.unc.edu/digccurr2007/papers/underwood_paper_6-3.pdf.

[12] Underwood, W., Hayslett, M., Isbell, S., Laib, S., Sherrill, S., & Underwood, M. (2009, October). *Advanced decision support for archival processing of presidential electronic records: Final scientific and technical report (ITTL/CSITD 09-05)*. Atlanta, GA: Georgia Tech Research Institute. Retrieved from: http://perpos.gtri.gatech.edu/publications/.

[13] Woods, K., & Brown, G. (2009). From imaging to access — effective preservation of legacy removable media. In W.G. LeFurgy (Ed), *Archiving 2009: Preservation strategies and imaging technologies for cultural heritage institutions and memory organizations: Final program and proceedings* (pp.213-218). Springfield, VA: Society for Imaging Science and Technology. retrieved from: http://www.imaging.org/IST/store/epub.cfm?abstrid=42914.

[14] 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. Retrieved from: http://dx.doi.org/10.1145/1998076.1998088.

[15] Xie, S. L. (2011). Building foundations for digital records forensics: A comparative study of the concept of reproduction in digital records management and digital forensics. *American Archivist*, 74(2), 576-599. Retrieved from: http://archivists.metapress.com/content/e088666710692t3k/.

About the Authors

Christopher (Cal) Lee is Associate Professor at the School of Information and Library Science at the University of North Carolina, Chapel Hill. His primary area of research is the long-term curation of digital collections. He is particularly interested in the professionalization of this work and the diffusion of existing tools and methods into professional practice. Lee edited and provided several chapters to I, Digital: Personal Collections in the Digital Era. He is Principal Investigator of the BitCurator project, which is developing and disseminating open-source digital forensics tools for use by archivists and librarians.

Matthew G. Kirschenbaum is Associate Professor in the Department of English at the University of Maryland and Associate Director of the Maryland Institute for Technology in the Humanities (MITH, an applied think-tank for the digital humanities). He is also an affiliated faculty member with the Human-Computer Interaction Lab at Maryland, and a member of the teaching faculty at the University of Virginia's Rare Book School. His first book, Mechanisms: New Media and the Forensic Imagination, was published by the MIT Press in 2008 and won the 2009 Richard J. Finneran Award from the Society for Textual Scholarship (STS), the 2009 George A. and Jean S. DeLong Prize from the Society for the History of Authorship, Reading, and Publishing (SHARP), and the 16th annual Prize for a First Book from the Modern Language Association (MLA). In 2010

he co-authored (with Richard Ovenden and Gabriela Redwine) Digital Forensics and Born-Digital Content in Cultural Heritage Collections, a report published by the Council on Library and Information Resources and recognized with a commendation from the Society of American Archivists. Kirschenbaum speaks and writes often on topics in the digital humanities and new media; his work has received coverage in the Atlantic, New York Times, National Public Radio, Wired, Boing Boing, Slashdot, and the Chronicle of Higher Education. He is a 2011 Guggenheim Fellow. See http://www.mkirschenbaum.net for more.

Alexandra Chassanoff is a Ph.D. candidate in the School of Information and Library Science at the University of North Carolina Chapel Hill. She is also the Project Manager for BitCurator. Alex's research interests center on how concepts of evidence, trustworthiness, and context are established for digitized archival materials. Prior to graduate school, Alex worked as a database analyst, IT consultant, and digital asset manager.

Porter Olsen Ph.D. candidate in the English Department at the University of Maryland. His research focuses on the intersections between postcolonial literature and digital cultures, with a particular interest in how both fields deploy virtual spaces as spaces of alterity. Before returning to graduate school, Porter worked as a product manager for a Linux distribution developer where he was a member of the United Linux initiative, an initiative designed to create a single Linux platform shared among distributors from Germany, Brazil, the U.S., and Japan.

Kam Woods is a Postdoctoral Research Associate in the School of Information and Library Science at the University of North Carolina at Chapel Hill. His research focuses on long-term digital preservation, data forensics, and file system analysis. He holds a Ph.D. in Computer Science from Indiana University Bloomington and a B.A. with a special major in Computer Science from Swarthmore College.

PRINTER-FRIENDLY FORMAT Return to Article