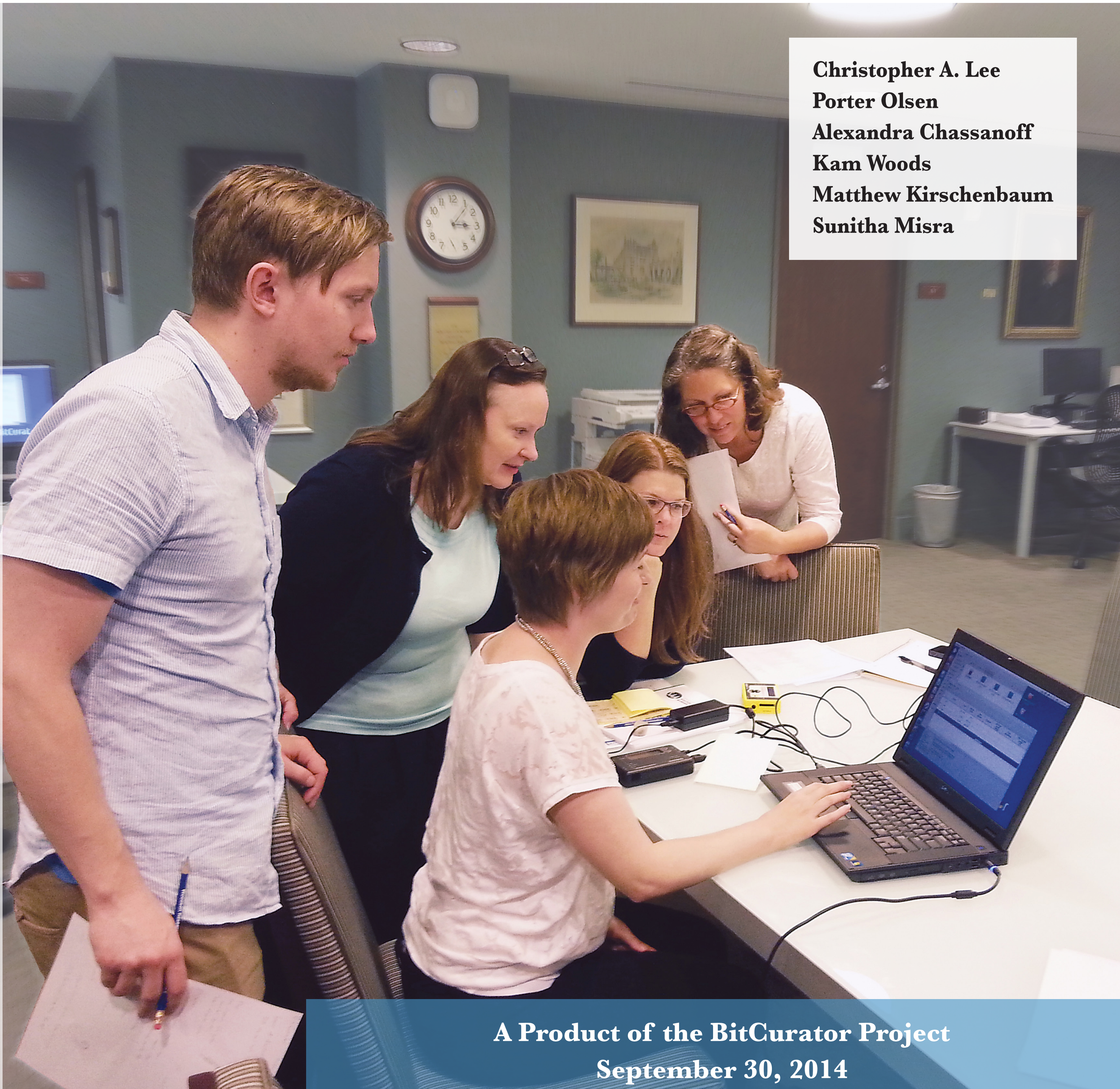
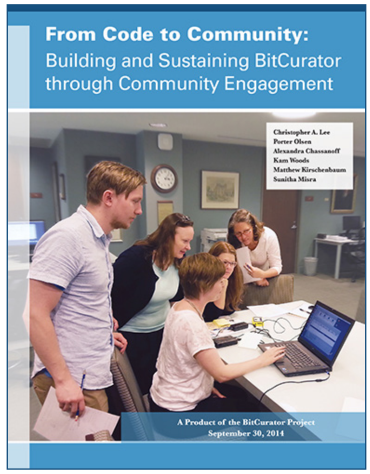
From Code to Community:
Building and Sustaining BitCurator through Community Engagement



Christopher A. Lee
Porter Olsen
Alexandra Chassanoff
Kam Woods
Matthew Kirschenbaum
Sunitha Misra

A Product of the BitCurator Project
September 30, 2014



From Code to Community:

Building and Sustaining BitCurator through Community Engagement

Christopher A. Lee
Porter Olsen
Alexandra Chassanoff
Kam Woods
Matthew Kirschenbaum
Sunitha Misra

A Product of the BitCurator Project
September 30, 2014

"You don't build open source software, you build open source communities."

-Paul Jones

1. Introduction

This paper examines efforts to develop and support a sustainable body of users of open-source digital forensics software within libraries, archives and museums (LAMs). It discusses motivations, challenges, and emerging strategies for the use of these technologies.

The BitCurator project ran from October 1, 2011 to September 30, 2014, through funding from the Andrew W. Mellon Foundation. The project was an effort to build, test, and analyze systems and software for incorporating digital forensics methods into the workflows of a variety of collecting institutions. It was led by the School of Information and Library Science (SILS) at the University of North Carolina, Chapel Hill, and the Maryland Institute for Technology in the Humanities (MITH) at the University of Maryland, and involved contributors from several other institutions. Two groups of external partners participated in this process: a Professional Expert Panel (PEP) of individuals who are at various stages of implementing digital forensics tools and methods, and a Development Advisory Group (DAG) of individuals who have significant experience with related software development activities.[1]

The BitCurator environment is a set of free and open-source tools designed specifically for LAMs. It can be installed as a Linux environment; run as a virtual machine (VM) on top of other operating systems (Windows, Mac, Unix/Linux); or run as individual software tools, packages, support scripts and documentation. Among its functionalities, the BitCurator environment allows individuals to create forensic disk images, perform data triage tasks, analyze and report on file systems, identify personal and sensitive information (such as social security numbers or credit card information), and enables the capture and exporting of technical metadata.

This paper is a product of the second phase of the BitCurator project (October 1, 2013 – September 29, 2014), which focused on expanding professional engagement and community outreach activities, along with ongoing development of software products.

1 Christopher A. Lee, Matthew Kirschenbaum, Alexandra Chassanoff, Porter Olsen, and Kam Woods, "BitCurator: Tools and Techniques for Digital Forensics in Collecting Institutions," *D-Lib Magazine* 18, no. 5/6 (May/June 2012), http://www.dlib.org/dlib/may12/lee/05lee.html.

2. Motivation and Related Work

LAMs have long served as venues for the preservation, management, description and provision of access to materials of continuing value. As the technologies for creating and using documentary artifacts have evolved, so too have the responsibilities and activities of LAMs. These institutions are now responsible for the curation of "born-digital" materials. Many LAMs already have a significant number of digital media (e.g. floppy disks, optical disks, hard drives) in their holdings, and such media continue to arrive as new acquisitions. Until quite recently, LAMs had few established approaches for dealing with born-digital materials stored on original media.

The incorporation of digital forensics tools and methods into LAM workflows has changed the landscape dramatically. Digital forensics can help to advance a variety of digital curation goals and activities, including: the creation of authentic copies of data on disks; reflection of the original order of materials; establishment of trustworthy chains of custody; discovery and exposure of associated contextual information; identification of sensitive information that should be filtered, redacted or masked in appropriate ways; redaction of specific bitstreams; and provision of access to the contents of disk images or directories of files.

Over the past five years, an increasing number of initiatives and institutions have advanced the application of forensics tools and methods in LAMs. This work has been documented in the BitCurator project's previous white paper.[2] The BitCurator project has contributed to these trends through the development, packaging, documentation and dissemination of an open-source software environment that has been specifically designed to meet the needs of LAMs processing and acquiring born-digital materials. Several elements of the BitCurator project were designed specifically to build capacity and ensure sustainability. The software is distributed under an open source license, so diverse constituencies can extend the tools at will. Members of the BitCurator team have developed and implemented a wide range of continuing professional education offerings – including a module for the Rare Book School (RBS), classes for the Digital Archives Specialist (DAS) curriculum of the Society of American Archivists (SAA),[3] components of the DigCCurr (Digital Curation Curriculum) Professional Institute, undergraduate/graduate-level classes at SILS, pre-conference workshops and a diversity of other one-time offerings (e.g. for the United Nations, U.S. Senate, National Library of Australia, National Library of New Zealand, Digital Curation Institute at the University of Toronto) – which have helped to cultivate a community of users. The BitCurator user electronic mailing list[4] provides a nexus for interested professionals to connect, and the project wiki[5] includes a substantial body of documentation to help people install and use the software.

In addition to the continuing software development work, the second phase (October 2013 to October 2014) of the project placed significant emphasis on professional engagement, community building and sustainability planning. This phase added a new role to the project, a Community Lead, who was responsible for actively promoting and supporting the incorporation of BitCurator tools into institutional practices. These outreach activities have been essential to the cultivation of a growing user community around the software. We have also developed and

2 Christopher A. Lee, Kam Woods, Matthew Kirschenbaum, and Alexandra Chassanoff, "From Bitstreams to Heritage: Putting Digital Forensics into Practice in Collecting Institutions" (September 30, 2013), http://www.bitcurator.net/wp-content/uploads/2013/11/From-Bitstream-to-Heritage-S.pdf.

3 Lee and Woods administered the SAA class 17 times.

4 https://groups.google.com/forum/#!forum/bitcurator-users

5 http://wiki.bitcurator.net

implemented the BitCurator Consortium (BCC) as a mechanism for stewarding and sustaining the software and the community supporting the project. The two issues (sustainability and community building) are closely related. The products of the BitCurator project will ultimately be sustainable if there are professionals working in a variety of institutions that value them, use them, and contribute to their ongoing development through evaluative feedback, bug reports and code revisions/enhancements.

3. Summary of BitCurator Project Activities Phase Two

The second phase of the BitCurator project focused on two main areas: (1) engaging with a user community and creating a model for sustainable open-source software supported by community members; and (2) continued iterative software development through enhancements to the BitCurator environment and refinements to the software.

Community Building and the BitCurator Project

From its inception, the BitCurator project has placed high priority on the long-term viability of the software, documentation, and workflows developed. The BitCurator team worked in both phases of the project to build an active, engaged user community around both software development and the application of digital forensics methods in the curation of digital collections. What follows is a review of community building efforts over the course of Phase 2 of the project.

Role of the Community Lead

The emphasis on community building during Phase 2 of the BitCurator project was reflected in the hiring of the BitCurator Community Lead, Porter Olsen. As the Community Lead, Olsen spearheaded the outreach and training efforts to help LAMs begin using the BitCurator environment and incorporate it into their institutional workflow practices. The Community Lead position was expected to "leverage both digital and in-person networks to bring BitCurator to a broad array of collecting institutions through 1) a series of hosted webinars, 2) on-site demonstrations, 3) evangelism at conferences and workshops within the archival community, and 4) coordination of the various means by which community members interact" (GitHub, the BitCurator project wiki, Google groups including bitcurator-users, email, and Twitter).[6]

Community Engagement Plan

One of the Community Lead's first tasks was to draft the Community Engagement Plan, which outlined key outreach activities and efforts that would take place during the second phase of the project. We identified key constituencies in the BitCurator user community, including: the BitCurator project team; collecting professionals working with born-digital collections in LAMs; the project funder (Andrew W. Mellon Foundation); the open source digital forensics community on whose work the BitCurator environment builds; the universities of North Carolina at Chapel Hill and Maryland; and providers of software that addressed other parts of digital curation workflows, such as Archivematica, ArchivesSpace, and Integrated Rule-Oriented Data System (iRODS). As an initial first step, identifying these constituencies and understanding their needs and expectations was an essential part of a wider effort to engage and build community.

Guided by the needs of the constituencies above, the Community Engagement Plan identified six key goals for Phase 2:

1. Raise awareness of BitCurator tools.
2. Educate LAM professionals in the motivations and methods for using open-source digital forensics tools and incorporating them into LAM workflows.

---

6 Christopher Lee and Matthew Kirschenbaum,"BitCurator: Tools for Digital Forensics Methods and Workflows in Real-World Collecting Institutions," Grant proposal to the Andrew W. Mellon Foundation, December 13, 2012.

3. Establish an active, self-sustaining BitCurator user group.
4. Develop and implement a process by which BitCurator users can contribute bug reports, request enhancements, and contribute code.
5. Provide opportunities for users to help build the BitCurator tool set through hackathons and other outreach efforts.
6. Lay the foundation for BitCurator's long-term sustainability by building a user base sufficiently invested in BitCurator's future that they will contribute time and resources to the project.

Members of the BitCurator Professional Experts Panel (PEP) provided valuable feedback in early drafts of the Community Engagement Plan.

Once community engagement activities were outlined, the Community Lead categorized tasks into four kinds of functional areas: Outreach, Communication, User Contact and Follow Up, and Education and Training. Table 1 summarizes tasks from the Community Engagement Plan and the functions they have supported. The tasks were carried out and supported by several members of the BitCurator project team. The Community Engagement Plan helped to identify potential gaps in the original project plan, items that were important to accomplish but were not already on our activities list.

Table 1: Community Engagement Plan tasks by function.

Task	Function
Identify early adopters and work with them to integrate BitCurator software into their workflows.	Outreach
Develop webinars ranging from an introduction to BitCurator to advanced user scenarios.	Outreach
Recruit individuals working at LAMs to the BitCurator users mailing list.	Outreach
Prepare BitCurator demonstration materials to be presented at conferences and other professional events.	Outreach
Actively blog through the project web site.[7]	Communication
Give presentations at LAM institutions and associated professional events.	Communication
Serve as an expert source for information regarding forensics tools in digital curation—primarily through the BitCurator wiki.[8]	Communication
Publish original research emerging out of the BitCurator project.	Communication
Collect names and contacts from users as BitCurator tools are adopted.	Contact and Follow up
Follow up with users to learn about successes or challenges with implementing the BitCurator tools.	Contact and Follow up
Integrate feedback from users into FAQs on the BitCurator wiki and in the Google Groups forum.	Contact and Follow up
Continue to refine and extend workshops and demos.	Education and Training
Develop online webinars addressing key functionality in the BitCurator tool set as they are completed.	Education and Training
Plan and conduct in-person site visits at selected institutions to provide hands-on training and integration of BitCurator into digital curation workflows.	Education and Training

7 http://bitcurator.net

8 Note that a valuable existing resource related to digital forensics in general (not specifically for LAMs) is the Forensics Wiki maintained by Simson Garfinkel: http://www.forensicswiki.org.

Site Visits

In order to train potential users and help speed adoption of the BitCurator environment, members of the BitCurator team conducted site visits at collecting institutions across the United States and Europe. These site visits grew in significance over the course of BitCurator Phase 2, as demand for them far exceeded our initial estimation. In addition to the primary goal of building a strong BitCurator user community, these site visits allowed us to test the BitCurator environment with real-world digital collections, which provided valuable user feedback. This section includes an overview of the general site visit structure and summaries of selected site visits.

Call for Proposals and Selection of Site Visits

Our initial plan called for up to five site visits during BitCurator Phase 2. We carefully considered which institutions to visit. While there were several compelling candidates at universities, we wanted to engage with a wider range of collecting institutions. In the winter of 2014, we issued a call for proposals (CFP). We asked submitting institutions to discuss their digital holdings, describe their progress to date with curation of born-digital materials, and identify the specific collection(s) they would be working with during the site visit. The goal was to work with institutions who were beyond the "getting started" phase and instead help those already processing born-digital content to integrate BitCurator into their workflows.

Figure 1: Professionals from universities, libraries, archives and other collecting institutions from across Sweden.

In the CFP we invited submitters to identify other geographically proximate organizations, and submit a joint proposal where appropriate. This strategy proved successful, allowing us to work with both the host institution and as many as three or four additional collecting institutions in the area. Most site visits involved multiple institutions.

We also requested that anyone submitting a proposal consider possibilities for cost sharing. This proved to be beneficial, essentially doubling the number of site visits we could conduct. Most of the site visit included some cost sharing contribution from the host. Financial benefit to the community building efforts aside, cost sharing also served as a powerful signal of institutional commitment. The need for tools such as those provided by the BitCurator environment was often driven by a recent acquisition that included important or valuable born-digital objects. The John Updike collection at Harvard was one notable and prominent example (see pages 27-28 for a discussion of using BitCurator to process born-digital materials from the collection).

The final element of the CFP was for the host institution to arrange a public outreach event such as a talk or a guest lecture. The integration of digital forensics tools and methods into digital preservation workflows requires a mixture of both technology and policy. To facilitate this, we reached out to administrators and other decision makers within the institution in addition to the professionals who would be using BitCurator on a day-to-day basis.

Table 2 below lists the dates and locations of the site visits conducted over Phase 2 of the BitCurator project.

Table 2: BitCurator Site Visit dates, locations, and institutions.

Date	Location	Participating Institutions
Sept. 19–20, 2013	State College, PA	Pennsylvania State University
Feb. 27, 2014	Evanston, IL	Northwestern University
Feb. 28, 2014	Urbana-Champaign, IL	University of Illinois, Urbana-Champaign
April 24–25, 2014	Princeton, NJ	Princeton University Institute for Advanced Study
May 7–9, 2014	Cambridge, MA	Harvard University
May 19–20, 2014	San Antonio, TX	University of Texas, San Antonio Trinity University
May 21–22, 2014	Austin, TX	University of Texas, Austin The Harry Ransom Center Texas State University Archives of the Episcopal Church Austin City Libraries
May 23, 2014	Houston, TX	Rice University
June 16, 2014	London, UK	The Tate Britain Kings College, London
June 17, 2014	London, UK	The British Library
June 20, 2014	London, UK	Archives and Records Association of the UK and Ireland
June 23–24, 2014	Stockholm, Sweden	Gothenburg University Library Umea University Library Swedish National Library Stockholm University Swedish National Museum Uppsala University Library Swedish National Archives Swedish Institute for Language and Folklore Stockholm City Archives Mid Sweden University Swedish Genealogical Association Swedish Civil Aviation Administration Center for Business History
June 25–26, 2014	The Hague, Netherlands	The Netherlands Coalition for Digital Preservation
June 27–30, 2014	Marbach, Germany	The German Literature Archive
July 1–July 2, 2014	Madison, WI	University of Wisconsin, Madison
July 9–10, 2014	Minneapolis, MN	University of Minnesota
June 16–17, 2014	New Haven, CT	Yale University Library Yale Center for British Art
June 17, 2014	Detroit, MI	Wayne State University Reuther Library
June 28–29, 2014	New York, NY	Metropolitan New York Library Council New York University Center for Jewish History
July 21, 2014	Washington, DC	U.S. Senate National Archives and Records Administration (NARA)
Sept. 8–9, 2014	Tucson, AZ	University of Arizona
Sept. 10–11, 2014	Phoenix, AZ	Arizona State University Arizona State Library, Archives and Public Records
Sept. 12, 2014	Flagstaff, AZ	Northern Arizona University
Sept. 15–16, 2014	Provo, UT	Brigham Young University Utah Valley University

Site Visit Structure

While the site visit structure evolved over the course of the year, it generally included three elements: 1) a day-long workshop, 2) a half-day practicum in which the skills obtained in the workshop were put into practice, and 3) a public outreach event. In the case of a one-day site visit, the workshop and practicum were combined into a single day and the public outreach event forgone.

Having multiple institutions at a workshop allowed us to emphasize the community-driven nature of our work. When Olsen conducted a site visit, he would begin by asking participants to introduce themselves, and ask those who were meeting someone in the room for the first time to raise their hands. In most cases, every hand would be raised. He would then ask those who were meeting someone from their own institution for the first time to leave their hand raised. It was often a surprise to the participants how many hands remained raised. This exercise emphasized the importance of building and maintaining communities within institutions, between local organizations, and ultimately with the wider BitCurator community. It also allowed individuals to discuss their various capacities (e.g. possession of digital forensics hardware, certain software expertise) and how those capacities could be shared among the group. There were frequently comments such as, "You have one of those? Can we come by and use it sometime?" The opportunity to see that others were encountering the same problems and that collectively they could accomplish more than they could on their own simultaneously reinforced the need for a community approach and helped lay the groundwork for building local community among BitCurator users.

The full day workshops began with an "Introduction to Digital Forensics" lecture. This lecture laid the ground work for the activities to follow in the workshop by defining digital forensics and its role in digital curation. This lecture introduced attendees to key digital forensics concepts such as checksums, hexadecimal representation of bitstreams, disk imaging, file system types and the technical metadata they record, and the means by which one can read data from legacy media.

With these foundational concepts laid out, the workshop progressed through the following tasks related to using BitCurator:

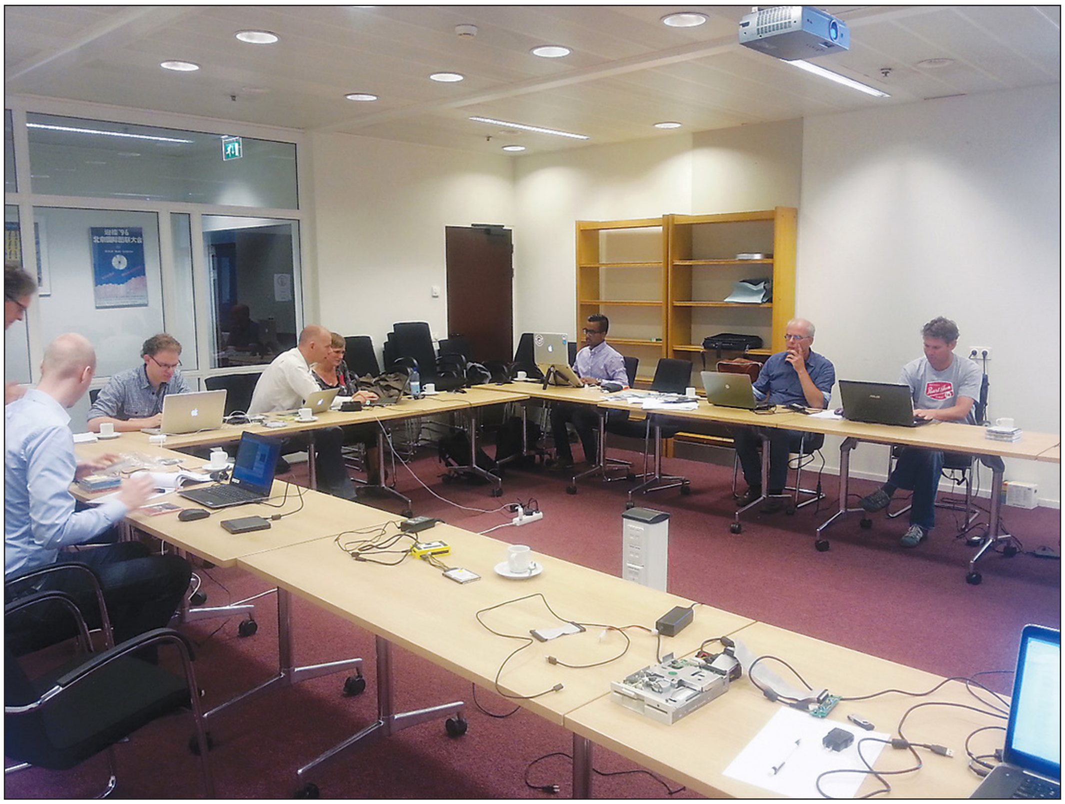
1. Installing BitCurator
2. Creating disk images with Guymager
3. Scanning for sensitive information with *bulk_extractor*
4. Generating DFXML technical metadata reports with *fiwalk*
5. Generating BitCurator reports
6. Using the BitCurator file access tool
7. Overview of other tools included in the BitCurator environment
8. Introduction to digital forensics and legacy disk access hardware

The workshops usually also introduced participants to hardware, including write blockers (used to prevent any changes to the data on a disk) and the FC5025 (a specialized device for reading 5.25-inch floppy disks with contemporary computers). Demonstrating the FC5025 interface was particularly useful because it underscored the importance of understanding file systems and that media of the same physical form factor may contain one of any number of potential file systems.

Because the workshops were participatory and hands-on, participants usually performed all of the tasks listed above. Due to the sometimes intimidating nature of digital forensics, there was particular value in "doing" rather than "seeing" in these workshops. This helped to dispel the notion that these tasks could only be performed by those with extensive technical training. By taking the time to ensure that each participant in the workshop—from administrators to processing archivists—could successfully complete each task, we demystified the concept of digital forensics and demonstrated straightforward applications of the tools.

The workshops also involved a practicum with the emphasis placed on working directly with the host institutions' digital collections. The aim of the practicum was to take the tools and concepts learned in the workshop and apply them to existing materials acquired by the institutions. We embraced the unpredictable nature of the practica because it emphasized the complexity of dealing with a diverse range of media types, file systems and file formats. Invariably, issues arose regarding how to access different types of media, the use of digital forensics hardware, challenges associated with identifying and manipulating various file systems, and the scope and limitations of the BitCurator software. Identifying and overcoming these challenges frequently gave attendees confidence that they would be able to perform these tasks once they were on their own. As a general rule, we tried to have each participant move through basic tasks several times. As with any learning process, the time to complete these tasks was reduced dramatically with each repetition.

Figure 2: Participants in the BitCurator practicum at the Netherlands National Archive.



The practicum also served as a key testing environment for BitCurator, providing us with the opportunity to assess user interaction and test a wide range of digital media. This often resulted in "exciting successes" (e.g. creating a disk image for the first time, quickly identifying sensitive information on a floppy disk using *bulk_extractor*) and "successful failures" (e.g. discovery of corrupted disks and hardware incompatibilities). Many of the features and refinements found in the BitCurator environment have their origins in these site visits, as well as feedback after performing hands-on tasks in other settings. These include the Rare Book School, the SAA DAS workshops and courses at SILS.

A key learning objective for the practicum, and the site visit overall, was for participants to understand how each of the tools in the BitCurator environment fit together within a larger digital curation workflow. Working through primary tasks associated with acquisition, processing, and analysis reinforced how these tools worked together – helping participants to see the BitCurator environment as a cohesive whole. This process of repetition also typically facilitated conversations about workflows and how BitCurator tools can fit into LAM practices. For those institutions that did not already have a workflow in place, the practicum gave them the chance to start asking the necessary policy questions related to this work.

Most BitCurator site visits culminated in a public outreach event, such as a talk or guest lecture. While the workshop and practicum focused on those individuals most likely to use BitCurator, these talks or lectures introduced BitCurator to a wider audience of students, university faculty, administrators, and an array of other interested parties. These events allowed us to articulate the challenges faced by LAMs when acquiring, processing, and analyzing born-digital materials. Addressing these challenges often requires active collaboration between several stakeholders—archivists and librarians, but also curators, administrators, researchers, and donors. How, for example, should the conversation between a donor and an institution change when a significant portion of the collection to be acquired will be received on digital media? How does understanding the concept of remanence—the residual data left on magnetic media after a user deletes a file—affect the framing of donor agreements or opportunities for research?[9] Identifying and answering these questions requires an in-depth understanding of the technical characteristics of digital media. Along with demonstrating the BitCurator environment, a two primary goals of public outreach events was to encourage institutions to start thinking through their born-digital workflows, and to prompt the critical "next step" conversations within the institution more broadly.

BitCurator Site Visits – Selected Highlights

Following each site visit, we created a Site Visit Trip Report which described institutions and individuals we worked with, provided any feedback received from participants, and summarized issues with BitCurator that emerged during the workshops. Trip reports ranged from relatively brief summaries to more formal documents several pages in length. The following site visit summaries are drawn from those more detailed trip reports. Unless noted otherwise, site visits were conducted by Porter Olsen in his role as the BitCurator Community Lead.

Northwestern University

In 2013, Northwestern University conducted a digital curation pilot in which they imaged the hard drive of a retired Northwestern professor. Paul Clough, a digitization systems librarian, invited us to visit Northwestern University Libraries to investigate how BitCurator could help the institution take further steps toward adopting digital forensics tools and methods.

The visit began with a live demonstration of BitCurator attended by a wide cross section of university library and archives staff. Participants from the library's special collections were interested in BitCurator's ability to work with audiovisual (AV) files. The AV preservationists recommended that we look into tools called *MediaInfo* and *FFProbe* for capturing AV file metadata, both of which have subsequently been integrated into the BitCurator environment. The university archivist raised a number of questions regarding access to deleted files and how one could ensure the destruction of digital files that were not meant to be donated. We discussed secure data deletion standards from the U.S. Department of Defense (DoD) and secure file deletion options. We tried to put this particular challenge in context by comparing it to the procedures an archive must undergo to deaccession paper records. Questions regarding secure deletion were prevalent throughout the remaining site visits.

After the demonstration, there was a visit to the office of Benn Joseph to see their digital curation pilot hardware and talk about how they might use BitCurator to further analyze the disk images they created in their pilot. Seeing the work they had done for the pilot emphasized the ongoing need for mature tools with detailed documentation that could assist institutions like Northwestern University Libraries as they explored the potential uses of digital forensics. It also reminded us that potential BitCurator users will range from those just beginning to process their born-digital collections, to institutions with dedicated digital archivists already using digital forensics hardware and software.

---

9 Gabriela Redwine, Megan Barnard, Kate Donovan, Erika Farr, Michael Forstrom, Will Hansen, Jeremy Leighton John, Nancy Kuhl, Seth Shaw, and Susan Thomas, *Born Digital: Guidance for Donors, Dealers, and Archival Repositories* (Washington, DC: Council on Library and Information Resources, 2013).

University of Illinois at Urbana-Champaign

The visit to the University of Illinois at Urbana-Champaign (UIUC) was an all-day training event organized by Tracy Popp, the Digital Preservation Coordinator at UIUC. There were nine people in attendance with roles that ranged from assistant university archivist to an IT representative responsible for maintaining the hardware necessary for the university's digital preservation work.

Participants brought their own laptops to the workshop which was held in a conference room. Popp brought hardware from her digital forensics lab, including a Digital Intelligence Forensic Recovery of Evidence (FRED)[10] workstation and a Kryoflux.[11] The value of having these items present in the workshop was quickly apparent—even though BitCurator can be run on a laptop, knowing how to use digital forensics hardware is a critical skill that most of the attendees did not already possess. Since Popp had already installed BitCurator on the FRED, participants were able to run the disk imaging process together as a group. After this initial demonstration of the BitCurator environment, the participants installed BitCurator on their laptops and began creating disk images. Popp had prepared disk images from media in her lab for the other participants to work with, but they preferred to create their own disk images in order to better understand the process.

As one of the first site visits, the workshop at UIUC significantly influenced the design and structure of future workshops. Three significant takeaways arose from the UIUC site visit and helped influence subsequent site visits:

1. Each workshop should begin with an introduction to basic digital forensics concepts. The technical backgrounds of participants at the UIUC site visit varied widely, from several novices in the group to Popp who had already been working with digital forensics hardware and software. It became clear that we needed to prepare BitCurator workshops with the full range of individuals who might attend in mind. This was consistent with our experience in other settings, including the SAA DAS classes and Rare Book School.
2. Participants strongly preferred creating their own disk images. Participants who were asked to bring their own media brought USB flash drives several gigabytes in size, which took 30-45 minutes (or longer) to image. To address this issue for future visits, we purchased ten 256MB USB flash drives for participants to use during the workshop. In addition to ensuring that the time to create their disk image would only be a few minutes, this also allowed us to prep the drives with information for the participants to discover using *bulk_extractor*.
3. Participants interested in integrating digital forensics tools and methods into their institution's workflows will need to understand how to use write blockers and access legacy data off floppy disks. In subsequent workshops, we included a section on digital forensics hardware, and brought those hardware devices (when possible) to give participants first-hand experience working with them.

Austin Texas Area Collecting Institutions

The site visit to the University of Texas at Austin was one of three visits across the state of Texas. The UT Austin visit was one of the first two-day workshops that included the second day practicum. The proposal from Austin was particularly strong because they had invited a number of collecting institutions to attend, including the Harry Ransom Center, Texas State University, Archives of the Episcopal Church, and the Austin City Libraries.

10 The FRED is a powerful specialized PC commonly found in digital forensics labs. It has built-in write blockers and the ability to swap hard drives in and out of the system without rebooting, among other features.

11 The Kryoflux is a custom floppy drive controller that allows access to a wide variety of legacy 3.5" and 5.25" floppy disk formats.

Figure 3: BitCurator workshop participants from Austin area collecting institutions.

As with earlier workshops, participants at UT Austin brought their own laptops. Installing the BitCurator virtual machine is a fairly straightforward process, but the wide range of laptops in workshops (including some which did not meet minimum hardware requirements) meant that significant time was spent on installation. One takeaway from this workshop was the recognition that it was helpful to have hosts provide lab access with the BitCurator environment preinstalled, or have attendees install BitCurator beforehand. This again reflects our experiences with other educational offerings, including the SAA DAS workshops and Rare Book School.

For the second-day practicum, most participants brought media from their institutions. Those who were not able to bring media were provided items from the UT Austin libraries. There was significant interest in accessing legacy media, so we setup a station with a variety of floppy disk controller cards. A floppy disk controller card acts as an intermediary between a legacy 3.5" or 5.25" floppy drive and a modern PC with USB support, as seen in Figure 4. The floppy drive data cable is connected to a controller card that understands how to read data from a wide variety of floppy disk sizes and formats. That information is then passed to the PC via a standard USB cable. Two commonly used floppy disk controller cards are the F5025, which can read data from 5.25" floppy disks, and the KryoFlux, which can read data from both 3.5" and 5.25" drives. Institutional licensing for the KryoFlux can cost several thousand dollars, so there is significant interest in a newer card called the SuperCard Pro, which – at the time of writing – retails for around $100 (US). The SuperCard Pro is the controller card being used in the photograph below.

Figure 4: Capturing data from legacy floppy disks with the Super Card Pro.

After the practicum, Olsen gave a guest lecture at the UT Austin School of Information on the need to preserve born-digital content and the use of BitCurator. Before leaving Austin, Olsen had an extended conversation with Jessica Meyerson of UT Austin libraries about the desire to support institutions in adapting existing workflows and systems for dealing with legacy media, rather than institutions having to invent solutions from scratch. Meyerson regarded the BitCurator environment as a means of breaking down the silos that exist within and between institutions.

Rice University

The site visit at Rice was unique in that it began with a public lecture and then moved to the hands-on workshop with a more select group of participants. Part of the reason for the departure from the normal schedule was so that more administrators and managers would be able to attend. Our hosts, Rice University archives and special collections, hoped that by introducing BitCurator to members of the libraries and archives more broadly, they would be able to generate stronger support for their efforts. This was one of the most poignant examples of the need to understand the complexities of preserving born-digital content at all institutional levels, not just the practitioners working directly with materials.

Because the IT department at Rice did not allow our hosts to install the BitCurator virtual machine, participants ran the BitCurator environment using the "live" bootable system image. This proved to be a challenge during the hands-on portion of the workshop. However, it also provided a chance to talk in detail about the different ways BitCurator can be run, including the appropriate use cases for each approach.

Figure 5: Archivists at the Woodson Research Center (Rice University) working with the BitCurator environment.

The site visit included a detailed discussion of how BitCurator can fit into Rice's existing workflows. They are trying to implement a repository that conforms to the Reference Model for an Open Archival Information System (OAIS),[12] and found the BitCurator reports to be helpful in the evaluation of disk images and file system contents and preparation of Archival Information Packages (AIPs).

Site Visits with Library and Information Science Schools as Hosts

Cal Lee conducted two site visits that were hosted by educators in library and information science schools. These visits were to Wayne State University in Detroit, Michigan, and the University of Wisconsin at Madison. Both involved a significant contingent of working

12 http://public.ccsds.org/publications/archive/650x0m2.pdf

professionals as well as students. At Wayne State, Kim Schroeder has been taking the lead on incorporating hands-on digital forensics activities into the curriculum, and Dorothea Salo has been doing the same at Wisconsin. In addition to the prepared talk and hands-on activities, Lee engaged in fruitful discussions with these educators about how best to further develop and implement curricula on the curation of born-digital materials.

Site Visits in the UK and Europe

After issuing the call for proposals, we received a number of requests for visits from institutions in the United Kingdom (UK), including: the British Library, the Archives and Records Association of the UK and Ireland (ARA), the University of Manchester, the National Archives of Scotland, and Kings College in conjunction with the Tate Museum. We also received site visit requests from Gothenburg and Umea University Libraries in Sweden, the Netherland Coalition for Digital Preservation, and the German Literature Archive in Marbach, Germany.

The Tate Britain and King's College

The first workshop we conducted in the UK was at the Tate Britain, which included participants from the Tate and King's College. There was a particularly strong contingent from the time-based media group at the Tate (8 out of the 11 participants). While most of their current collections are not born-digital, they predict that born-digital content will be a significant part of future acquisitions, and want to build expertise in dealing with these materials in advance of that change.

Figure 6: A view of the Tate Britain in London.

One issue discussed during this site visit was internationalization. At issue is both support for local languages in Ubuntu (the operating system upon which the BitCurator VM is built) and how the various digital forensics tools handle non-ASCII character sets. As the BitCurator community grows beyond the United States, adapting these tools to function effectively for users in other countries and different languages will be critical. We recently learned that BitCurator has drawn the attention of information schools in Japan, further emphasizing the need for additional internationalization efforts.[13] The day concluded with a discussion between Olsen and Mark Hedges regarding the various efforts being pursued as part of the PERICLES project and what relation they might have to BitCurator.[14]

The British Library

The second site visit in the UK was to the British Library (BL), hosted by BitCurator Professional Experts Panel member Jeremy Leighton John. While at the British Library, Olsen gave a talk on the role of digital forensics in LAMs as part of the 21st Century Curatorship lecture series. Once again, this provided a chance to speak to a broad cross section of the institution's employees.

Figure 7: Jeremy Leighton John in the electronic manuscripts lab at the British Library running BitCurator on a workstation.

After the talk, Olsen visited Jeremy's digital forensics lab - one of the first of its kind in a collecting institution. There, Jeremy demonstrated how he had integrated BitCurator into his digital forensics practices.

Due largely to Jeremy's efforts, the BL already has an established track record of using digital forensics tools and methods. The workshop introduced others within the BL to the significance of these practices, and the means by which they could be extended to the rest of the library. In addition to BL employees, there were also participants from the National Library of Scotland and Manchester University Libraries. In both cases, participants were excited about the possibilities BitCurator offered and planned to take what they had learned back to their respective institutions.

Archives and Records Association of the UK and Ireland

Simon Wilson, the senior archivist at Hull University, invited Olsen to address the UK and Ireland Archives and Records Association Section for Archives and Technology. This event was hosted at University College London. It was a talk and demo of BitCurator and part of the ARA Section for Archives and Technology Annual Meeting. While there was not an opportunity for hands-on work at this event, it provided an opportunity for further outreach activities.

Stockholm Sweden

We were originally approached by Gothenburg and Umea University libraries to conduct a BitCurator site visit in Gothenburg, Sweden. As word spread that we would be conducting a workshop on digital forensics and BitCurator, a number of Sweden's other collecting institutions asked to attend, necessitating a move to the Royal Library of Sweden in Stockholm. The event was the largest of the site visits held in Europe and included more than a dozen collecting institutions from across the country (see Figure 1).

13 See: http://current.ndl.go.jp/node/19353
14 http://pericles-project.eu/

Figure 8: A site visit at the Swedish National Archive.

The second day of the site visit took place at the Swedish National Archives. This provided a classroom setting (shown in Figure 8) and access to the archive's digital collections. The head of technology for the Swedish National Archive welcomed us, but was unable to stay for the workshop. He shared a stack of 5.25" disks and jokingly told us that as payment for using the room he wanted the group to create disk images of the disks for him. Twenty minutes later, imaging of the floppy disks was complete. In addition to repaying the host's hospitality, imaging the floppy disks gave a number of the participants a chance to work with legacy media and the FC5025 floppy drive controller.

The Hague Netherlands (Netherlands Coalition for Digital Preservation)

Marcel Ras, the program manager for the Netherlands Coalition for Digital Preservation (NCDD), invited us to give a workshop for NCDD members (shown in Figure 2). This was the most technically advanced group of all the participants during the two weeks conducting BitCurator workshops in Europe. There was much discussion of tool automation and how the output from *fiwalk* and the *BitCurator Reporting Tool* could be incorporated into institutional repositories. One of the organizers of the visit stressed that part of the value of learning about digital forensics is that it changes the way one thinks about digital objects in collections, a sentiment that we made a point to emphasize in the site visits and other educational events.

Because of the technical skills in this group, during the hands-on practicum they quickly moved beyond the basics and began pushing BitCurator to see its limits. Specifically, a couple of participants started imaging their cell phones, with varying success.

Site Visits: Conclusion and Lessons Learned

An important lesson learned from conducting BitCurator site visits was the need for agility and flexibility. At each stage in the process—from initial outreach, to workshop design, to the call for proposals, to the site visits themselves—we continually needed to adjust to the needs and requirements of our hosts and site visit participants. Had we held rigid ideas about the role that site visits would play in community building efforts, we would have failed to capitalize on these opportunities to engage with these potential BitCurator users. When the call for proposals generated far more interest than we had initially anticipated, Cal Lee worked with the Mellon Foundation to ensure that we had sufficient funds to accommodate the high numbers of requests. Porter Olsen, the Community Lead and the individual conducting most of the site visits, needed to learn from the early visits and build a site visit curriculum that met the needs of the host institutions as well as the community building goals of the BitCurator team. Kam Woods, the BitCurator Technical Lead, needed to make time available when BitCurator team members were conducting site visits so he could quickly address the technical issues that emerged during the hands-on workshops. Through this flexibility we were able to provide detailed, hands-on training to over two dozen institutions instead of just the handful initially planned.

In addition to the need for flexibility, three other key lessons were learned over the course of BitCurator site visits:

1. The importance of community buy-in: By inviting host institutions to participate in cost sharing, we substantially expanded the number of site visits we were able to conduct. Cost sharing also engendered a sense of shared commitment.
2. Focus on local communities: Another lesson that emerged from our call for proposals was the importance of local communities of practice. Being able to gather people together in a room who are in the same institution or geographical area and have them discuss how to address the challenges they face by working together went a long way towards building community. The emphasis on local community opened up possibilities for participating in larger national or international community around the BitCurator tools. We have had similar experience when running other events for staff of specific institutions (e.g. United Nations, U.S. Senate, National Library of Australia, National Library of New Zealand).
3. Invite many stakeholders: Digital curation activities require support and contributions from many different levels at an institution. While a processing archivist may want to apply digital forensics tools, they will need institutional support for and appreciation of the complexities of this work.

Software Development: Responding and Adapting to Community Needs

At the start of Phase 2, many of the tools and interfaces under development by the BitCurator team were in beta; functional, but not production-ready. Development of the BitCurator tools and the BitCurator environment during Phase 2 focused on improving performance and stability of existing software (including the BitCurator Reporting Tools), in addition to introducing and refining several new tools. Areas of focus were driven by community interactions and new testing regimes implemented in the project. This included feedback provided in the BitCurator users group, hand-on experience with users during workshops, and a broader range of synthetic and real-world test corpora used by the teams at SILS and MITH. The following sections discuss improvements and additions to software produced by the BitCurator team; addition and modifications of third-party tools; alterations and improvements to the environment; testing; and documentation.

BitCurator Tools

In keeping with our commitment to agile development methods, the software tools developed by the BitCurator team evolved significantly between 2013 and 2014. Much of this development was conducted in response to observing and speaking with members of the community conducting trials with the existing toolset. In several cases, entirely new tools were developed during Phase 2 to address gaps in functionality when working with physical media and disk images. Each of the tools developed by the BitCurator project team, as they exist in the 1.0 release, is described below.

BitCurator Reporting Tool

The BitCurator Reporting Tool was originally developed as a command-line Python script allowing users to automate the process of extracting file system metadata from disk images (by calling the *fiwalk* tool integrated into The Sleuth Kit), annotate features of interest (including potentially private and sensitive information) identified in disk images by Simson Garfinkel's *bulk_extractor*, and produce text, PDF, and Excel reports to support analysis and curation of born-digital materials.

In July 2013, the BitCurator team released a preliminary graphical user interface (GUI) for this tool. The GUI included individual tabs for *fiwalk*, the *bulk_extractor* output annotation scripts, and the BitCurator report generation module. While this interface simplified the process of interacting with these relatively complex tools, it still required relatively detailed knowledge on the part of the user regarding the order in which the individual tools should be run in order to generate the final reports. In addition, the tool provided limited feedback to the user during long-running processes.

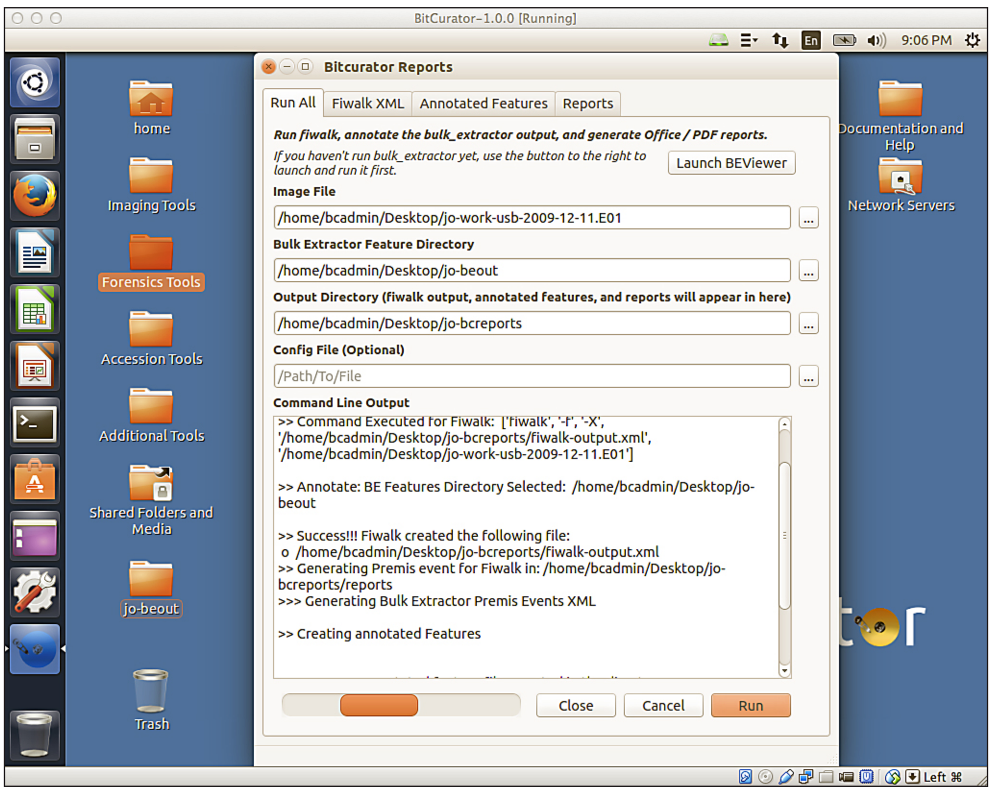
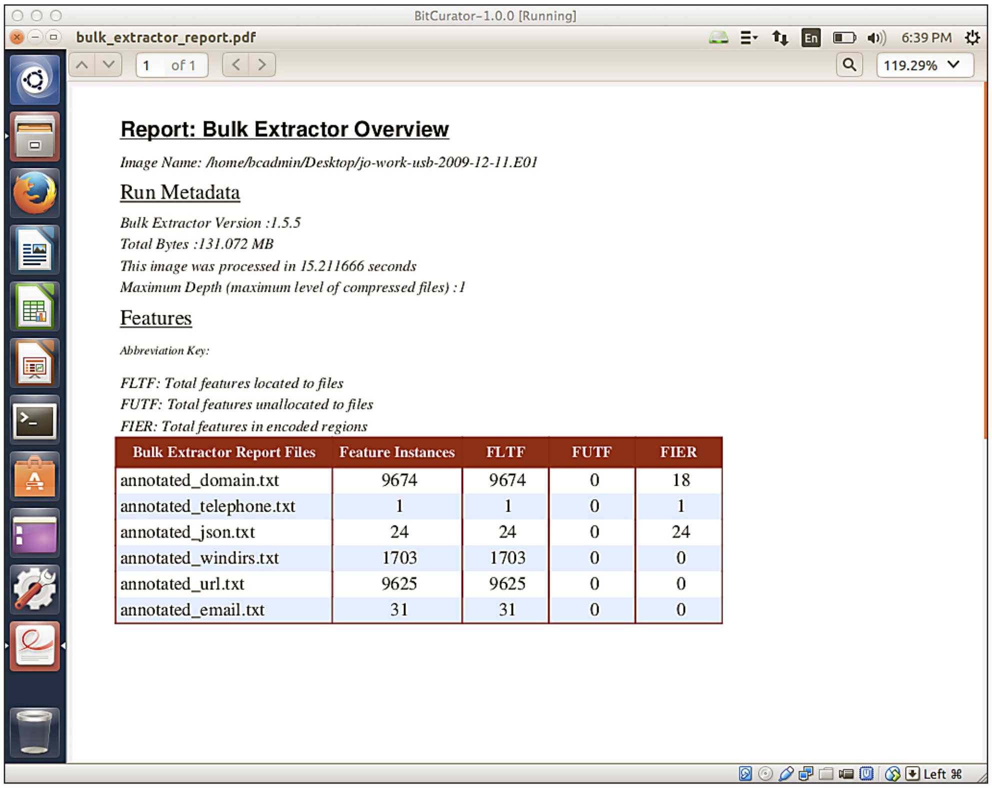


Figure 9: The BitCurator Reporting Tool processing a disk image.

Further development of the BitCurator Reporting Tool during Phase 2 focused on addressing these issues. A *Run All* tab was added to the existing interface to clarify the intended workflow: simplifying access to the graphical front-end to *bulk_extractor* (*BEViewer*), and allowing users to run all subsequent report processing stages in a single window. The original Python code, which ran serially in a single processing thread, was rewritten to allow individual data processing tasks to run in separate threads from the main interface and provide real-time feedback on progress.

Configuring the output of the BitCurator Reporting Tool was also simplified, allowing users to select or deselect PDF and Excel reports by editing a simple text file (the default version of this configuration file is found in the directory */etc/bitcurator/bc_report_config.txt* in the BitCurator environment).

Each of the reports produced by the Reporting Tool have been substantially modified, both in response to user comments and as the digital forensics software libraries used by BitCurator to generate them have expanded in scope. Early efforts during the project focused primarily on



Report: Bulk Extractor Overview

Image Name: /home/bcadmin/Desktop/jo-work-usb-2009-12-11.E01

Run Metadata

Bulk Extractor Version :1.5.5
Total Bytes :131.072 MB
This image was processed in 15.211666 seconds
Maximum Depth (maximum level of compressed files) :1

Features

Abbreviation Key:

FLTF: Total features located to files
FUTF: Total features unallocated to files
FIER: Total features in encoded regions

Bulk Extractor Report Files	Feature Instances	FLTF	FUTF	FIER
annotated_domain.txt	9674	9674	0	18
annotated_telephone.txt	1	1	0	1
annotated_json.txt	24	24	0	24
annotated_windirs.txt	1703	1703	0	0
annotated_url.txt	9625	9625	0	0
annotated_email.txt	31	31	0	0

Figure 10: Overview of a bulk_extractor report, including processing metadata.

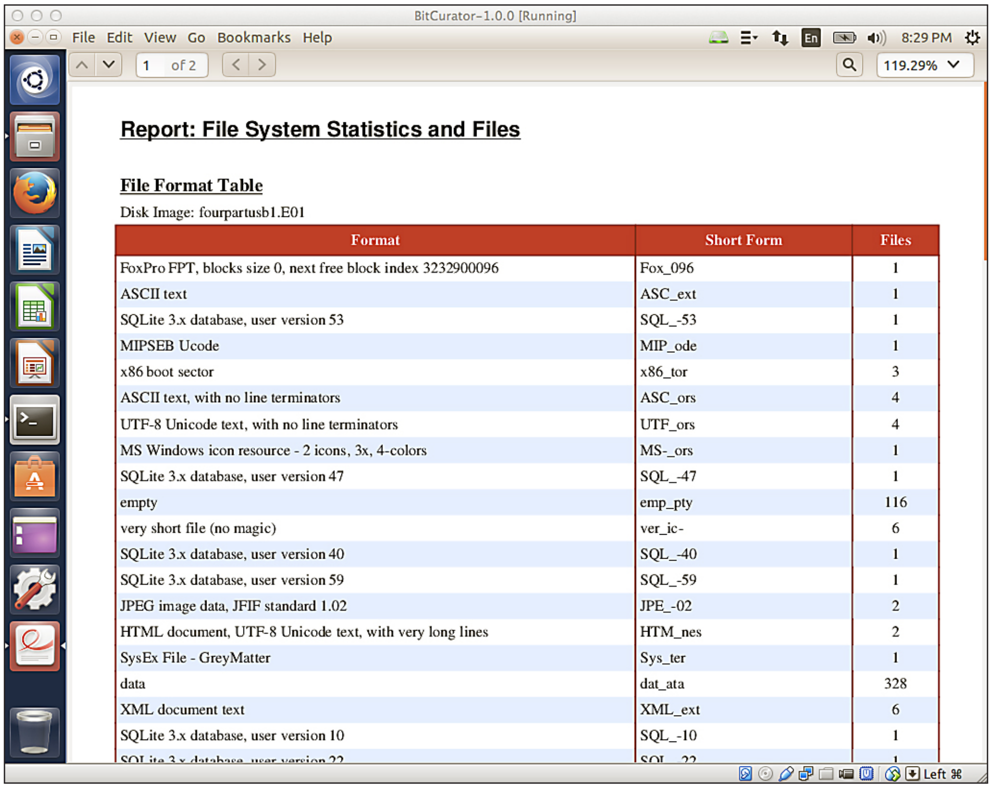
generating static, informational reports (generated as PDFs) that could be used to analyze the structure and contents of disk images. Many of these, such as the "Bulk Extractor Overview" report (seen in Figure 10) have been revised to provide an "at-a-glance" review of the disk image being processed, the amount of time required to process that image, the size of the disk image file, and other relevant capture metadata (in addition to a high-level report of the features identified by *bulk_extractor*).

These reports capture a wide range of information that may be useful in making decisions about the eventual disposition of the data being analyzed; identifying items that may incorporate private or potentially sensitive information and flagging them for possible redaction; performing identification of both file systems and file format types; and providing other more general statistics about the contents.

This level of detail – the granularity at which the digital objects are processed – may not always be actionable in a preservation context. For example, the BitCurator Reporting Tool relies on *libmagic* output incorporated into the output to produce PDF reports on file type distribution within disk images. The *libmagic* tool uses a database of "magic" to identify files based on specific sequences of bytes appearing within the bitstream. In typical implementations, this database includes more than 11,000 sequences – far more unique identifiers than the set of official MIME types[15] or the approximately 1100 identifiers (at the time of writing) in the PRONOM database.[16] In the BitCurator environment, this discrepancy (and the different use cases associated with each of these forms of identification) is addressed both by providing a high-granularity report (as seen in Figure 11) based on the *libmagic* output, and allowing users to configure *fiwalk* to perform file type identification with other tools using a plugin mechanism (the DOMEX Gateway Interface).

Another major focus of development efforts was improving users' ability to interact directly with disk images within the BitCurator environment. While it was possible for users of earlier iterations of the BitCurator environment to mount forensically packaged disk images, examine

15 http://www.iana.org/assignments/media-types/media-types.xhtml
16 http://apps.nationalarchives.gov.uk/PRONOM/Default.aspx



Report: File System Statistics and Files

File Format Table

Disk Image: fourpartusb1.E01

Format	Short Form	Files
FoxPro FPT, blocks size 0, next free block index 3232900096	Fox_096	1
ASCII text	ASC_ext	1
SQLite 3.x database, user version 53	SQL_-53	1
MIPSEB Ucode	MIP_ode	1
x86 boot sector	x86_tor	3
ASCII text, with no line terminators	ASC_ors	4
UTF-8 Unicode text, with no line terminators	UTF_ors	4
MS Windows icon resource - 2 icons, 3x, 4-colors	MS-_ors	1
SQLite 3.x database, user version 47	SQL_-47	1
empty	emp_pty	116
very short file (no magic)	ver_ic-	6
SQLite 3.x database, user version 40	SQL_-40	1
SQLite 3.x database, user version 59	SQL_-59	1
JPEG image data, JFIF standard 1.02	JPE_-02	2
HTML document, UTF-8 Unicode text, with very long lines	HTM_nes	2
SysEx File - GreyMatter	Sys_ter	1
data	dat_ata	328
XML document text	XML_ext	6
SQLite 3.x database, user version 10	SQL_-10	1

Figure 11: Default format identification of files using libmagic in the BitCurator Reporting Tool.

their contents, and copy out items, this required a series of relatively complex command-line interactions. During Phase 2, we introduced a simple contextual menu allowing users to right-click on disk images and mount the contained file system(s) read-only.

One of the benefits of generating forensically packaged disk images – in formats such as Encase image file format[17] and Advanced Forensic Format (AFF)[18] – is the availability of software libraries that can quickly identify deleted and unallocated items within the bitstream (items that cannot be accessed in a mounted disk image). In order to address the need to simply identify and export these items, the BitCurator team developed a tool – first introduced in BitCurator 0.7.0 on February 2, 2014 – to export selections of files from a given disk image. This tool was originally implemented as an additional tab in the BitCurator Reporting interface. It constructed a hierarchical display of the file system that could be browsed by the user, and allowed individual files to be exported to a pre-selected directory.

BitCurator Disk Image Access Tool

While the implementation described above was sufficient for simple tasks ("export all files from this disk image"), it was rudimentary and provided few additional features compared to commercial offerings, such as AccessData's FTK Imager.[19] To address this, we deprecated the original interface and introduced a new tool in July 2014, called BitCurator Disk Image Access (DIA). The DIA interface allows users to load both raw and forensically packaged disk images, select and export files based on status (allocated and unallocated), and view disk image capture metadata. It provides a color-coordinated view of volumes available within the disk image (in green), directories (in bold), allocated files (in black), and unallocated or deleted items (in red). A sample view of the DIA interface operating on a disk image with partitions containing four distinct file systems can be seen in Figure 12.

17 http://www.forensicswiki.org/wiki/Encase_image_file_format
18 http://www.forensicswiki.org/wiki/AFF
19 http://www.accessdata.com/support/product-downloads

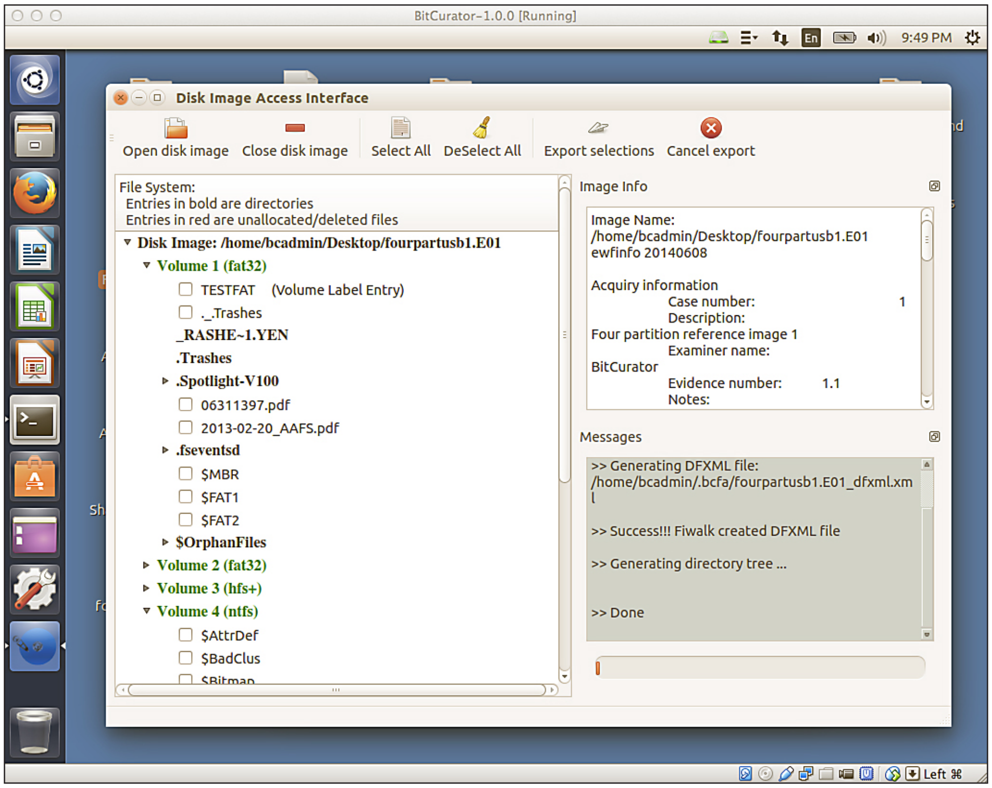


Figure 12: The BitCurator Disk Image Access Interface.

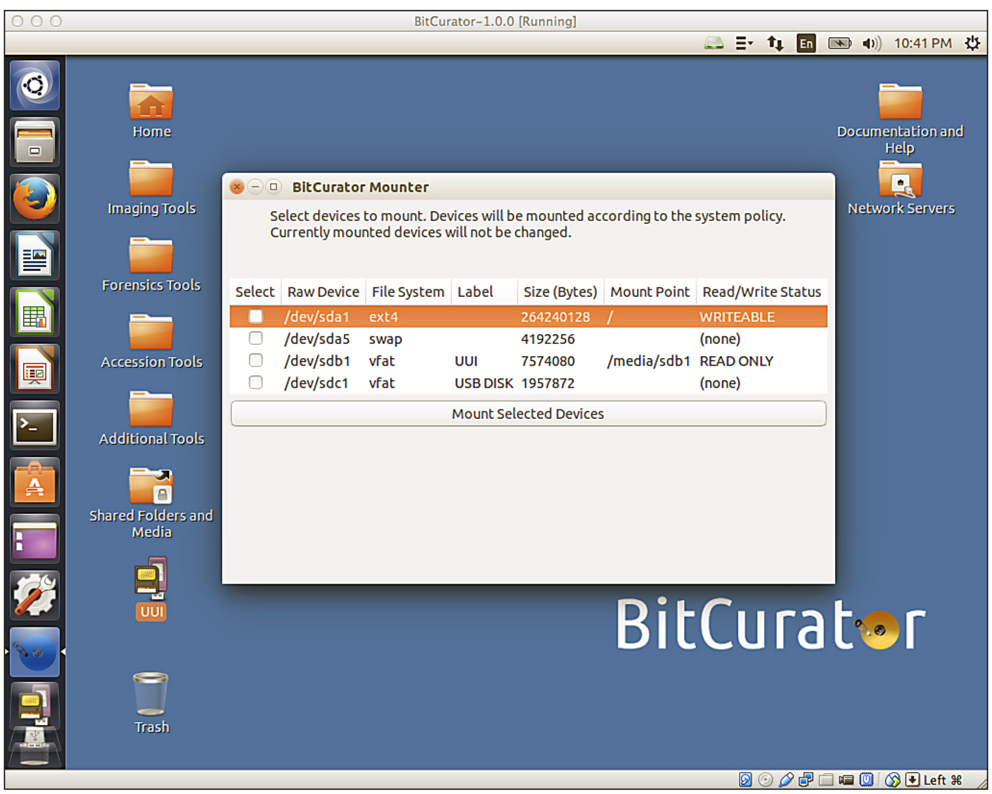


Figure 13: BitCurator Mounter showing two attached USB devices.

Every version of the BitCurator environment has relied on tools and scripts from other forensics distributions to facilitate ease of interaction with physical media and the disk images extracted from those media. Over time, changes to the Linux environments on which these systems have been based have rendered certain aspects of these tools inoperable. In May 2014, we reconstructed the BitCurator environment using a new "Long Term Service" release of Ubuntu (14.04, or "Trusty Tahr"). With this release, the open source scripts originally developed by a third party[20] and modified to provide read-only mounting of physical media (and allow users to reenable read/write access) were rewritten as standalone tools to address these breakages.

The first of these tools is the BitCurator Mounter (see Figure 13). Essentially a clone of the "Mounter" script written by John Lehr using *yad* (Yet Another Dialogue), the BitCurator Mounter is a Qt application written in Python. While Ubuntu includes a built-in utility for mounting physical devices (*udisks*), the BitCurator Mounter provides users with an alternate view of mounted and available physical devices, identifying those devices that are currently mounted read-only and read/write in a simplified dialogue. In this image, the "WRITEABLE" disk is the virtual drive on which the VM is running. The "READ-ONLY" disk, mounted according to the current mount policy, can be seen as an icon on the desktop to the left of the BitCurator Mounter interface.

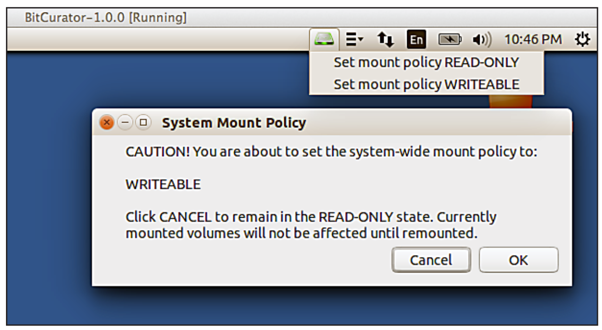


Figure 14: The BitCurator write policy AppIndicator.

In tandem with development of the BitCurator Mounter, the mounting policy scripts were rewritten by the BitCurator team to operate correctly in Ubuntu 14.04. The Ubuntu "AppIndicator" application was developed to support this policy (see Figure 14).

Third-Party Tools

During Phase 2, many additional tools and utilities were added to (and updated within) the BitCurator environment to support common actions performed on born-digital materials. A full listing of these tools can be found on the Software page on the project wiki.[21] Notable additions during this period included current versions of the File Information Tool Set (FITS)[22] suite, the Library of Congress' BagIt Library[23] and Bagger GUI, and the Clam Antivirus engine[24] (along with scripts to execute virus scans against forensically-packaged disk images).

BitCurator Environment

The start of Phase 2 coincided with our fourth major beta release of the BitCurator environment (0.4.0). This was an important milestone for the project; it was the first release allowing users of the BitCurator Reporting Tool to run the entire reporting sequence over raw and forensically packaged disk images as a single stage. This eliminated the need to run potentially confusing, intermediate processing steps – such as annotating features of interest located by the *bulk_extractor* tool with associated filenames within the disk image – by using a dedicated panel within the GUI interface to run all steps in a single pass.

Twelve major releases of the environment were issued during the next six months, largely focusing on integrating the improvements described in the previous sections. As noted in the previous section, we transitioned the environment to the new "Long Term Service" release of

20 John Lehr developed the mounting script and a range of other utilities used in the CAINE forensics environment, https://code.google.com/p/linuxsleuthing/.
21 http://wiki.bitcurator.net/index.php?title=Software
22 http://projects.iq.harvard.edu/fits
23 https://github.com/LibraryOfCongress/bagit-java
24 http://www.clamav.net/index.html

Ubuntu (14.04) in May 2014. In addition to ensuring the ongoing viability of the environment design, moving to Ubuntu 14.04 provided us with a simpler way to access up-to-date software libraries needed to compile and package current versions of many of the required forensics tools. It also allowed us to address the issue of "cruft" that had accumulated in the environment (old versions of tools, configuration scripts, temporary files, legacy kernels, and other dead artifacts). While relatively mundane, this is a problem that plagues many of the existing law-enforcement-specific environments available today. We have put significant effort into ensuring that every tool, configuration file, and support script in the BitCurator environment is functional and up-to-date at the time of release.

Testing

Ensuring proper functionality of a software environment requires a consistent and well-documented testing regime. In addition to ongoing monitoring of the unit and regressions tests for software developed by the BitCurator team (and tracked via our GitHub repository), we established a number of straight-forward mechanisms to allow team members to report software failures, bugs, and behavioral glitches without having to write formal bug reports.

The team developed a shared repository of small (generally less than 256MB) disk images covering a range of different file systems, media formats, and known-problem edge cases. These were tracked using a shared spreadsheet. Core tools such as the BitCurator Reporting Tool and the BitCurator Disk Image Access tool were tested against these disk images, adding new images as they became available or were developed by the team to address newly discovered problems.

The majority of this testing took place during the last few months of Phase 2 and was performed by Kyle Bickoff, the Graduate Research Assistant on the BitCurator Project at the University of Maryland. While the testing itself was still relatively time-consuming, using a well-documented corpus of disk images reduced the "time to fix" for existing and newly introduced bugs significantly. A sample of the spreadsheet used to track these issues can be seen in Figure 15. Note that the accompanying text for the failure cases is brief but precise. We found that these kinds of reports worked well for our small but high-availability team; any failure that could not be immediately repeated (or traced to an obvious software bug) could generally be addressed in an impromptu meeting.

Figure 15: Sample of a testing spreadsheet used by members of the BitCurator team.

fx Name

	A	B	C	D	E	F
1	Name	BitCurator Version	Media Type	File System	Export 'Deleted'	Export 'Intact Files (aka Not-Deleted files)'
2	*Example*	*Release #*	*Example 3.5" Floppy*	*Example NTFS*	*Example success/not properly exported/no result, or not even readable with BC reporting tool*	
78	Test 76 - (Test 16 #4)	0.9.21	Zip100	Unknown (fiwalk fails)	FAIL - Loads into DIAI, but directory tree cannot be accessed	FAIL - Loads into DIAI, but directory tree canno
79	Test 77 - (Test 17 #4)	0.9.21	Zip100	Unknown (fiwalk fails)	FAIL - Loads into DIAI, but directory tree cannot be accessed	FAIL - Loads into DIAI, but directory tree canno
80	Test 78 - (Test 18 #4)	0.9.21	Zip100	FAT16	Success	Success
81	Test 79 - (Test 20 #4)	0.9.21	Zip100	FAT16	Success	Success
82	Test 80 - (Test 3 #5)	0.9.22 (RC1)	3.5" Floppy	FAT12	Success	Success
83	Test 81 - (Test 5 #5)	0.9.22 (RC1)	USB	FAT32	FAIL - Loads into DIAI, but directory tree cannot be accessed	FAIL - Loads into DIAI, but directory tree canno
84	Test 82 - (Test 6 #5)	0.9.22 (RC1)	USB	Unreadable	FAIL - Loads into DIAI, but directory tree cannot be accessed	FAIL - Loads into DIAI, but directory tree canno
85	Test 83 - (Test 7 #5)	0.9.22 (RC1)	DVD-RW	ISO9660	Success	Success
86	Test 84 - (Test 8 #5)	0.9.22 (RC1)	Iomega ZIP250	FAT16	Success	Success
87	Test 85 - (Test 11 #5)	0.9.22 (RC1)	CD-R	ISO9660	Success	Success
88	Test 86 - (Test 12 #5)	0.9.22 (RC1)	CD-R	ISO9660	Success	Success
89	Test 87 - (Test 14 #6)	0.9.22 (RC1)	USB	FAT32	FAIL - Loads into DIAI, but directory tree cannot be accessed	FAIL - Loads into DIAI, but directory tree canno
90	Test 88 - (Test 22 #5)	0.9.22 (RC1)	Zip100	FAT16	Success	Success
91	Test 89 - (Test 25 #5)	0.9.22 (RC1)	USB	FAT16	Success	Success
92	Test 90 - (Test 26 #5)	0.9.22 (RC1)	USB	NTFS	Success	Success
93	Test 91 - (Test 27 #5)	0.9.22 (RC1)	USB		Success	Success
94	Test 92 - (Test 2 #5)	0.9.22 (RC1)	3.5" Floppy	FAT12	Success	Success
95	Test 93 - (Test 16 #5)	0.9.22 (RC1)	Zip100	Unknown (fiwalk fails)	FAIL - Loads into DIAI, but directory tree cannot be accessed	FAIL - Loads into DIAI, but directory tree canno
96	Test 94 - (Test 17 #5)	0.9.22 (RC1)	Zip100	Unknown (fiwalk fails)	FAIL - Loads into DIAI, but directory tree cannot be accessed	FAIL - Loads into DIAI, but directory tree canno
97	Test 95 - (Test 18 #5)	0.9.22 (RC1)	Zip100	FAT16	Success	Success
98	Test 96 - (Test 20 #5)	0.9.22 (RC1)	Zip100	FAT16	Success	Success
99	Test 97 - (Test 3 #6)	"1.0"	3.5" Floppy	FAT12	Success	Success
100	Test 98 - (Test 5 #6)	"1.0"	USB	FAT32	FAIL - Loads into DIAI, but directory tree cannot be accessed	FAIL - Loads into DIAI, but directory tree canno

An important lesson learned in preparing our test corpora was that it was generally faster to create synthetic data to replicate problems encountered by community members than to solicit "real-world" disk images from LAMs. In the majority of cases, institutional policy, pre-existing donor agreements, security concerns, and other distribution restrictions prevented institutions from sharing their data with us. Over the course of the project, we prioritized creation of surrogate disk images that we believed addressed the most common use cases.

Documentation

A primary goal of the BitCurator project has been to provide documentation and support mechanisms to ensure that LAM professionals know how to use the software, when and where to apply it, and have at least one trusted resource to consult when they encounter problems.

Cornerstones of our strategy have included a consistently updated "Quick Start Guide,"[25] creation and curation of a public wiki with additional written documentation, a series of video walkthroughs, and direct communication with the BitCurator community via the google group user forum, professional events, and site visits.

The wiki has evolved significantly over the course of the project. In each edit, we have attempted to ensure it meets three basic criteria:

1. Simplicity: The wiki is first and foremost a technical resource for users who may have limited prior experience with these tools, and time constraints on learning what they do and how to use them. Over time, we have modified the language in every entry of the wiki to ensure it is both easy to understand and limits assumptions about prior knowledge.
2. Consistency: Wiki sites work best when they have many eyes reviewing pages. As our team is relatively small, we have used automated tools and periodic reviews to ensure there is no contradictory, unreadable, or incomplete information on the site.
3. Clarity: Our primary documentation appears on the front page of the wiki, organized as a sample workflow that reflects the intended use cases of the software.

Finally, documentation within the environment itself was expanded during Phase 2. The documentation folder which appears on the BitCurator desktop now includes the BitCurator Quickstart, user manuals for *bulk_extractor*, *BEViewer*, and *sdhash*, and the Digital Forensics XML schema and tag library. When appropriate, we have also made tools within the environment "self-documenting," including improved descriptive labeling in the BitCurator Reporting Tool and tooltips for interface elements in the BitCurator Disk Image Access tool.

Software Testing, Evaluation and Tool Review

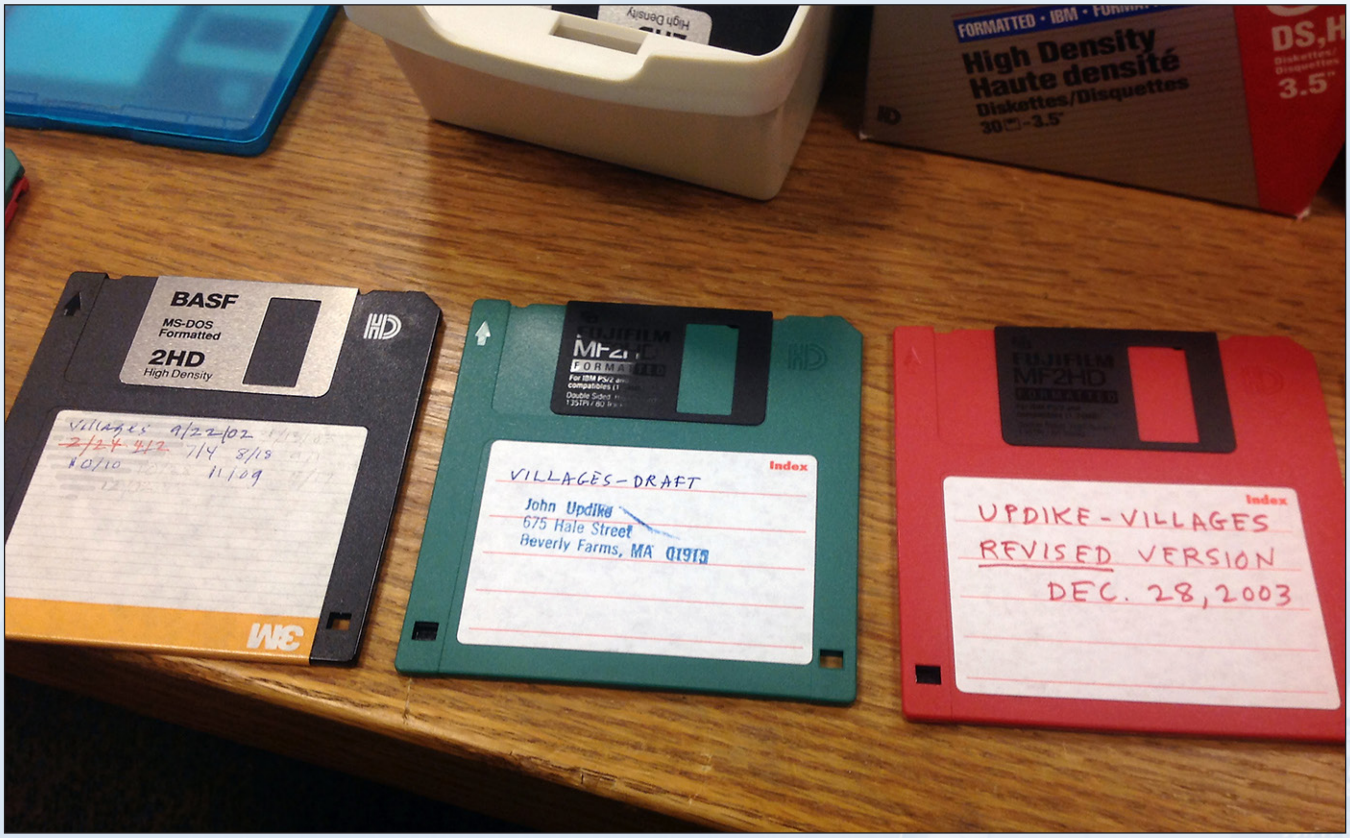
The iterative software development approach used for BitCurator involved continuous testing and refinement to ensure functionality in a variety of settings. Efforts taken toward this approach included the following: (1) assembling and testing a corpus of forensic materials; (2) creating the "BitCurator-in-a-Box" program for members of the LAM community to test out the BitCurator environment; (3) eliciting feedback from Professional Experts Panel (PEP) and Development Advisory Group (DAG) members as they performed task-based evaluations of the BitCurator environment; and (4) organizing the "BitCurator Clinic" – an on-site workshop for North Carolina Triangle-area archivists and librarians who could bring materials from their own institutions and work closely with BitCurator.

Corpus Development

A major challenge to improving the consistency and coverage of private data handling across collecting institutions is the lack of shared corpora on which to test software designed to identify such data. It is unlikely that such a corpus could be created from existing collections materials (or unprocessed materials within backlogs) due to existing donor agreements, legal guidelines, and institutional mandates. Although it is possible to construct large corpora of data from private materials inadvertently or maliciously released onto the Web, this can be ethically problematic to further disseminate, and such corpora will not generally include the complex data structures and interrelationships found on media originally belonging to individual users and organizations.

(Continued on Page Twenty-Eight)

25 http://wiki.bitcurator.net/downloads/BitCurator-Quickstart.pdf



Rabbit's Bits Run

Figure 16: Three 3.5" diskettes from the John Updike Collection acquired by Harvard University's Houghton Library.

On a drizzly Cambridge morning last May, Porter Olsen and I from the BitCurator team found ourselves in a basement room with personnel from Harvard's Houghton Library, including digital archivist Melanie Wisner and Leslie Morris, one of the manuscripts curators. The star of the show, though, was Mr. John Updike, whose born-digital remains arrived unceremoniously in a big brown box.

Updike as much as anyone has earned the right to the title of dean of American letters in the second half of the twentieth century. He began using a Wang word processor in 1983, the very year his biographer Adam Begley identifies as the height of his literary career. After a decade or so, he moved over to IBM PCs and the word processing software associated with them, first Lotus Ami Pro and later Microsoft Word. Updike never got rid of his typewriters, or indeed his pen though: instead, the computers took their place as part of his workflow, instruments of composition for letters, short stories, and essays while novels and poems he continued to draft longhand.

The Houghton acquired the Updike papers in 2009, and though the processing is ongoing the collection is open for research with an excellent finding aid (I had worked with the papers back in the fall). In the biography, Adam Begley had described Updike's papers as perhaps "the last all-paper collection of its kind." But this is not quite right: there are also about fifty 3.5" high-density IBM-compatible diskettes, of which 38 appear to have content on them; a half dozen 5.25" floppies which are installation disks for Lotus Ami Pro; and a dozen or so CD-ROMs. There are no hard drives or complete computers, nor do any of the diskettes from the Wang era appear to have survived.

But the born-digital materials that do survive as part of the collection are part of the author's manuscript record, and until there is a sustained scholarly investigation of them we cannot know what, if anything, they might contain in the way of drafts and other materials that could shed light on some aspect of Updike's literary career. For me this was an opportunity to get a firsthand look at the digital life of a writer I was also researching for my book on the literary history of word processing; for the BitCurator team, it was an opportunity to work with cultural heritage materials of paramount importance.

The Houghton had prepared for us a Mac Mini with 16 GB of RAM and Virtual Box and the BitCurator virtual machine preinstalled, as well as a USB 3.5" drive suitable for the high-density IBM-formatted diskettes we knew we would be working with. Over the course of several hours (that included discussion and instruction as well) we imaged a dozen of the disks without incident; one initially manifested bad sectors but corrected itself after a repeat of the imaging process. For each image we did a quick, initial inspection using *bulk_extractor* and the BitCurator reporting tools.

There was no smoking gun "LostNovel.doc", nor had I really expected to find one. But there was a palpable sense of accomplishment in the room, as the librarians present realized that *this is doable*. As Porter and Melanie worked hands-on with the diskettes, conversations sparked around issues like file naming conventions, directory structures, and what to represent in a finding aid, as well as, of course, strategies for researcher access. All of this was very gratifying to see. For my part, I did a preliminary sort and arrangement based on the disks' labels, and then manually write-protected each disk using its plastic slider mechanism. Sitting in a basement room of the Houghton and manually write-protecting John Updike's computer disks was not something I ever anticipated doing in a scholarly career!

My prior experience with the papers taught me that Updike was frugal, and reused all manner of material, typing or writing on the backs of drafts, or even envelopes and receipts and other people's correspondence. Certainly his digital working habits appear consistent. His practice was apparently to store multiple versions of a file on the disk, overwriting previous ones with new ones and notating the date on the disk's label after crossing out what was written previously. For some novels, like *Villages, Terrorist,* and the *Widows of Eastwick* (sequel to the more famous *Witches*, which was done before his word processing days) there are multiple diskettes with relevant material; others contain several dozen shorter pieces such as stories or reviews. One is marked C:\FAMILY\JOHN and would seem to contain personal material. Of course we do not know what he may or may not have happened on his hard drives. Moreover, he was in the habit of producing multiple hard copy typescripts in the course of working on a book, and then annotating and revising these by hand. For some lucky researcher, there will be an

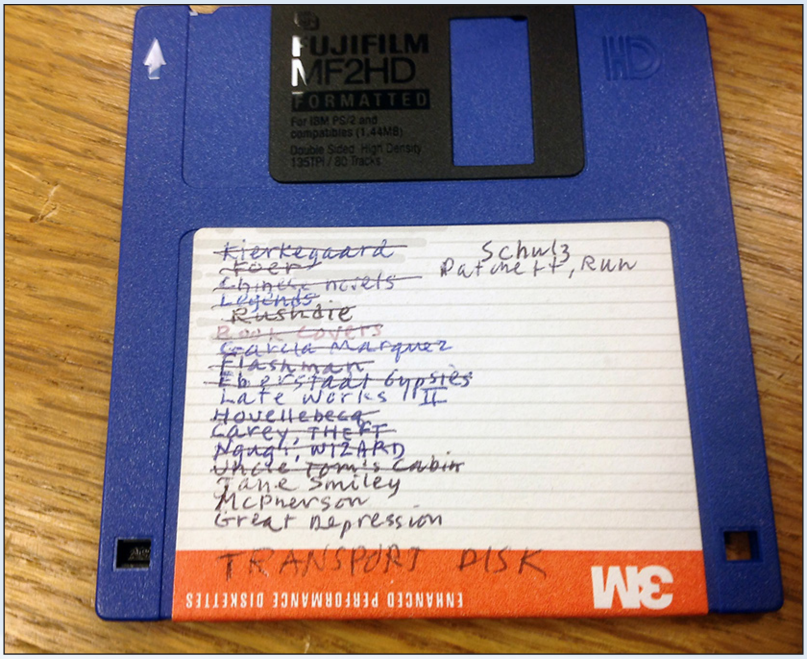


Figure 17: Close up view of a 3.5" diskette from the John Updike collection.

interesting challenge in triangulating between a.) the hard copy typescripts or print-outs in the physical papers, which are usually dated, b.) the dates on the labels of the diskettes, and c.) the modified, accessed and changed/created (MAC) times on the diskettes and the contents of their digital files, including deleted TMP files. Whether major insights into Updike's creative life are thus revealed or not, this is paradigmatic of what literary textual scholarship is going to look like in the coming years.

The Houghton staff indicated that they need to think through the file management and policy issues, but they are prepared to move ahead rapidly with the imaging and wish to be responsive to future requests from researchers.

– *Matthew Kirschenbaum*

(Continued from Page Twenty-Five)

The BitCurator Project has constructed a test corpus to replicate common data analysis and triage workflow steps. This consists of a non-public corpus of disk images extracted from fixed and removable media identified as containing data likely to have archival value (or requiring long-term preservation) by researchers and practitioners from the project's two advisory groups. In the first year of the project, we requested data from the ten project advisors on the PEP, and nine on the DAG. We received data in the form of raw and forensically packaged disk images from the City of Vancouver Archives, Duke University, the National Institute for Standards and Technology (NIST), the National Library of Australia, and the University of North Carolina (Wilson Library). The transfer was based on a data transfer agreement in which the project team agreed to use the data only for purposes of research and testing within the context of the project.

We have subsequently added to this corpus approximately ten years of disk images from retired workstations and legacy external media provided by iBiblio at the University of North Carolina at Chapel Hill. Additionally, we have included approximately 100,000 government documents in common office file formats crawled from the Web for the purposes of sampling document metadata and content. The corpus includes approximately 7.5TB of data, with coverage of major disk formats including FAT16 and FAT32, NTFS, HFS and HFS+, Ext2/3/4, and various double- and high-density floppy images.

BitCurator in a Box

BitCurator-in-a-Box (BCiaB) was a pilot study designed to help institutions to easily get started in using the BitCurator environment to process born-digital materials and to provide us feedback on their experience. Recruited participants came from 12 different institutions, including but not limited to: academic libraries, historical museums, state archives, and private foundation archives. BCiaB began rollout during the first three months of Phase 2 (October – December 2013) and completed in March 2014.

Figure 18: BitCurator in a Box.

Each institution in the program received the following in a Pelican-box evaluation kit:

- Printed documentation detailing each task in the evaluation, including: installation, disk image creation, locating personally identifiable information (PII), and generating the BitCurator digital forensics reports
- A flash drive containing the BitCurator environment as a VirtualBox virtual machine and an ISO image that could be installed as its own Linux environment
- A hardware write-blocker – as needed – for the appropriate media storage format used in their evaluation (the hardware was loaned and then returned)
- A link to an online response form/questionnaire where participants were asked to provide detailed feedback about their experiences working with the BitCurator environment.

As part of the pilot evaluation, participants were asked to perform six tasks using the BitCurator environment: (1) install and launch BitCurator as a virtual machine; (2) create a disk image using Guymager; (3) run bulk_extractor and use the BEViewer GUI to examine features of interest for potentially private and sensitive information; (4) run fiwalk on the disk image; (5) generate annotated bulk extractor output; and (6) generate BitCurator-produced digital forensics reports. Participants were instructed to use the BitCurator wiki documentation for guidance as they performed each task. Following the completion of each task, participants were asked to answer six questions describing their experience with that task.

Once all participants had completed their evaluation, the project team compiled and analyzed responses. Feedback was then categorized into the following areas: BitCurator Environment, Documentation, General Environment, and Downloads. The most common issues reported concerned installation of VirtualBox and launching the BitCurator environment. Two participants noted they were unable to install VirtualBox due to insufficient administrative privileges. Another participant mentioned having to call a team member to determine how to create a new virtual machine.

The project team was able to address the majority of suggestions/bugs encountered through enhancement or clarification in our documentation. For example, we created new screencasts to demonstrate detailed instructions on how to run each tool. Another change was the addition of highlighted "tips" throughout the Quick Start Guide, so that common issues (typically related to administrative privileges) could be addressed. Discrepancies between the software environment and the documentation were also fixed.

Elicitation of Feedback from Advisory Group Members

Another significant activity was the elicitation of feedback from both PEP and DAG members on current documentation and software usability. The project team asked advisory board members to evaluate the BitCurator environment while performing the six basic tasks used in the BCinaB pilot evaluation.

Three members (two from the PEP and one from the DAG) completed all six tasks and provided useful feedback in their evaluations. Most feedback centered on making the wiki documentation more usable and accurate for participants installing the BitCurator environment. For example, participants noted minor discrepancies between screenshots in the wiki documentation and the current release environment. Another participant mentioned that further explanation about BitCurator environment tools would be helpful.

We also asked participants to describe which types of documentation they used in completion of the tasks. Online instructions available on the BitCurator wiki were mentioned by all three participants, while two mentioned the tutorial videos available on the BitCurator YouTube channel. When asked where they went for help after encountering a problem, all three respondents mentioned the wiki. One respondent also mentioned the BitCurator users list.

Further insight was provided when they were asked about the most useful place(s) they went for help. One respondent felt that the BitCurator users mailing list was the most useful because of the human component. Another respondent speculated he/she would probably find the mailing list most useful, but had not been aware of its existence and thus had only consulted online wiki documentation. The third participant felt that the online wiki documentation was the most useful because it was "quick." Two of the three mentioned a preference for using wiki instructions rather than screencasts.

Digital Lives Research Workshop

On September 11-12, 2014, Cal Lee, Jeremey Leighton John (British Library) and Susan Thomas (Bodleian Library) ran a workshop at the British Library called "Applying Forensics to Preserving the Past: Current Activities and Future Possibilities." This was supported through funds from the Digital Lives project, headed by John. The event was held at the British Library Conference Centre. The focus was on application of digital forensics to acquisition, processing and preservation of materials in libraries, archives and museums. Participants reported on current activities, discuss gaps and opportunities and advance collective action. The objectives were to feature innovate application of digital forensics in LAMs; report on the current state of the art; identify challenges, gaps, and opportunities for further collaboration; and articulate recommendations for future activity. There were 15 participants, from the UK, US and Denmark. The majority of participants submitted one-page papers and then gave short presentations based on the papers. Most focused on current practices and challenges within their respective institutions. There was a set of breakout discussions that metadata, user access, review for sensitive data, and documenting workflows. While a written report from the workshop is still forthcoming, there is already evidence of the event's potential impact. For example, one set of proposed action items from a breakout discussion was to undertake the enhancement of the Community Owned digital Preservation Tool (COPTR) Registry[26] to better reflect digital forensics tools and to supplement COPTR with documentation of institutional workflows. Lee has followed up with the leadership of the COPTR initiative on these points, and they have been very well received.

Figure 19: Kam Woods discussing legacy magnetic media at the BitCurator Clinic.

BitCurator Clinic

As part of our efforts to gather real-world scenarios for testing the BitCurator environment, the project team at UNC Chapel Hill organized the "BitCurator Clinic." The event was a day-long workshop for archivists and librarians from area institutions to learn more about using BitCurator for processing born-digital materials. Eight people attended and brought materials from their respective institutions, which included: the Wilson Library at UNC Chapel Hill, State Archives of North Carolina, and the North Carolina State University.

Participants were instructed to bring born-digital materials from their respective institutions, preferably in the form of already-created disk images to cut down on the time spent imaging.

26 http://coptr.digipres.org/

This event raised a number of interesting use cases. One case involved imaging of a small partition on a 3TB drive. The participant, who was command-line-savvy, said at one point "I didn't realize I could actually use *dd* to image just this partition," and was later happy to see how Guymager could be used to produce new types of images from existing images. This participant said, of that overall experience "That, right there, made this entire clinic a win for us." We also conducted live disk imaging of floppies from an archival collection which included both images and EXIF metadata. Finally, we demonstrated how to use specific tools in the BitCurator environment, including a tool facilitating deduplication.

4. Future Work

Future work can be considered in two broad categories: (1) maintenance and cultivation of the software products and associated user community, and (2) further development to support additional aspects of digital curation workflows. In the first category, we will discuss the BitCurator Consortium. In the second category, we will discuss the forthcoming BitCurator Access project.

BitCurator Consortium

An essential player in the future application of digital forensics tools in LAMs will be the BitCurator Consortium (BCC).[27] The BCC will serve as the host of and center of administrative and user support for the BitCurator environment. This will include the software developed as part of the BitCurator project and, when appropriate and feasible, software developed by (1) additional follow-on research and development projects (including BitCurator Access, described below), and (2) members of the BitCurator user community. Governance decisions will be driven by BCC members, which will include an Executive Council along with targeted committees. The BCC operates as an affiliated community of the Educopia Institute, a non-profit organization that advances cultural, scientific, and scholarly institutions by catalyzing networks and collaborative communities to facilitate collective impact. As of the time this paper was completed, ten institutions have joined the BCC as either charter or general members, and there are many more that are in the process of joining.

BitCurator Access

BitCurator Access (October 1, 2014 – September 30, 2014), led by the School of Information and Library Science at the University of North Carolina, Chapel Hill (SILS), will build on the work of the BitCurator project (Phase 1 and Phase 2). It will focus on simplifying and improving access to the content of disk images held in born-digital collections.

The ultimate goal of the proposed project is to enable new forms of research and discovery based on born-digital materials. There has been a significant shift in recent years toward the adoption of digital forensics tools and methods by libraries, archives and museums (LAMs). This process has been facilitated by the BitCurator project. Many LAMs across the globe are using these tools and methods to generate disk images, extract metadata to support ongoing preservation tasks, and store the resulting data in dedicated servers or shared network spaces. However, there is currently limited support for provision of access to the contents of the disk images or associated metadata. BitCurator Access is a project designed to address this pressing need.

The project will pursue three main approaches to providing access to the data: setting up a server that holds the full disk images and then lets end users dynamically walk the directory tree and access the individual folders and files; exporting both the files from the disk and the metadata using forensics tools (in the form of DFXML) and loading the files and DFXML into a more traditional collection access environment so that people can search and navigate the metadata; and access to disk image content through emulation.

Also closely associated with the above access scenarios is redaction. The BitCurator environment allows LAMs to identify sensitive content, summarize the results and map the

27 http://www.bitcurator.net/bitcurator-consortium/

identified features to specific files and folders. There are currently some rudimentary tools and methods for redaction of data from disks, directories and individual files, but they are still very immature. BitCurator Access will develop tools to redact files, file system metadata, and targeted bitstreams within disks or directories. This will be offered both through the command line and GUIs integrated into the BitCurator environment.

A design priority is to ensure close integration between the existing functionality of the BitCurator environment and the software developed by the BitCurator Access project. This will allow institutions to run the access tools on the same machine (or virtual machine) as the one they are using for the initial processing. Institutions may instead elect to run these tools in separate, dedicated server environments in order to better manage and allocate their resources.

5. Conclusions

The application of digital forensics tools and methods in LAMs is still a relatively new phenomenon, but there has been a substantial shift toward adoption in recent years, facilitated (in large part) by the work of the BitCurator project. We have been able to develop a software environment that is responsive to various pressing needs in LAMs, and we have interacted with hundreds of professionals working in those settings. There are seven particularly notable factors that account for progress on BitCurator development and adoption:

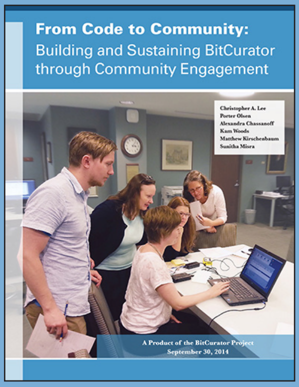
1. Clearly defining the scope of the project so that it addresses a set of identified needs within LAMs;
2. Agile development with frequent releases that are responsive to specific requirements or problems elicited from working professionals;
3. Community engagement and outreach, including educational offerings and site visits;
4. Strong collaborative relationships between the project partners at SILS and MITH;
5. Dissemination of the software as free and open-source;
6. Development and public distribution of documentation and guidance resources; and
7. Planning for sustainability early in the process, so that the BitCurator Consortium could be fully formed and active before the end of the project.

Success in the curation of born-digital materials is constituted by ongoing programs of activity, learning, adaptation and improvement, rather than a single, discrete product. Similarly, a successful open-source software initiative is one that can continually enroll human and technical resources to perpetuate and evolve the software. We hope that this paper provides lessons and insights that can help others to take on similar endeavors.

6. Acknowledgements

BitCurator was supported by two grants (for phases 1 and 2) from the Andrew W. Mellon Foundation. Members of the BitCurator project team have been Kyle Bickoff, Alexandra Chassanoff, Matthew Kirschenbaum, Christopher (Cal) Lee, Sunitha Misra, Porter Olsen, Amanda Visconti, and Kam Woods. We would like to acknowledge the members of the BitCurator Development Advisory Group (DAG) and Professional Expert Panel (PEP) for the for their valuable support and input: Bradley Daigle, Erika Farr, Geoffrey Brown (2011-2012), Barbara Guttman, Jeremy Leighton John, Leslie Johnston, Jennie Levine Knies, Jerome McDonough, Mark Matienzo, Courtney Mumma, Naomi Nelson, Erin O'Meara, Michael Olson, David Pearson, Gabriela Redwine, Doug Reside, Seth Shaw, Susan Thomas, William Underwood, and Peter Van Garderen (2011-2012).

We would also like to thank the numerous LAM professionals who have provided us individual feedback on the project's products, approaches and strategies. Finally, we would like to acknowledge Katherine Skinner of the Educopia Institute for her instrumental role in bringing the BitCurator Consortium into being.



From Code to Community: Building and Sustaining BitCurator through Community Engagement

Christopher A. Lee, Porter Olsen, Alexandra Chassanoff, Kam Woods, Matthew Kirschenbaum, and Sunitha Misra