# Boutique to Production – Insights from Stanford's Born-Digital / Forensics Labs

***Michael Glenn Olson; Digital Library Systems and Services; Stanford University; U.S.A.***

## Abstract

*Beginning in 2014 Stanford University Libraries has made dramatic changes in how it operates its Born-Digital / Forensics Labs (BDFLs). The goal has been a shift from a boutique lab operated by a single digital archivist to a production service that can handle digital collection materials and patron requests in bulk. This paper will outline Stanford's new practices and will include a discussion of how we budget, staff our labs, track our work, configure our hardware / software workstations, and generate statistics to support the preservation of our research collections.*

## Introduction

In 2009 Stanford University Libraries created its first Born-Digital / Forensics Lab to begin preserving its existing and newly acquired holdings of born-digital acquisitions. These include complete computer systems acquired from faculty and donors, collections containing a wide variety and quantity of computer storage media. With a single digital archivist on staff and an increasing backlog that is greater than 30,000 pieces of media we needed to innovate how we operate our lab by making changes to how we budget, train staff, configure our equipment, track our work and generate statistics to document our value.

The Goal of these recent changes has been to dramatically increase the volume of media we are able to preserve and reduce the cost per item of computer media preserved. This paper will describe these changes in how Stanford operates its Born-Digital Forensics Labs and to share this information with the Library, Archive and Museum community.

## History of our Labs

Our lab began in 2009 with the purchase of a two Forensic Recovery of Evidence Devices (FREDs). The first was a tower workstation with built a built-in write blocker suite and the second was a laptop for doing acquisition of born-digital content off site. By 2013 we were now operating two labs with three FREDs and eight custom built workstations specifically designed for recovering data from floppy disks, optical media and Iomega Zip disks. We distinguish these capture workstations from our FREDs as they were designed specifically to recover data from legacy computer media and not for processing born-digital content. Our digital archivist custom built these workstations around the only modern motherboard we could find in 2012 that still contained a floppy disk controller chip. This allowed us to connect 3 ½" and 5 ¼" floppy drives directly to our workstations. We also rounded out our lab hardware with suites of fixed and portable write blockers, multiple Catweasels and a Kyroflux [1].

Both of our labs were staffed by a single digital archivist. Our digital archivist did the capture and processing of high priority born-digital collections and trained and oversaw the work of project archivists responsible for preserving and describing collections that contained born-digital materials.

## Digitization as a Service

Beginning in fiscal year 2014 Stanford Libraries began an effort to structure and market our labs as a service. The goal was to provide our subject and public service librarians with a clearly defined process for how they could get materials digitized in any of our labs irrespective of the physical format of the materials. A subject or public service librarian requesting digitization of a video, a map, or recovery of data from an obsolete piece of computer media should follow the same process and all digitization work in any of our labs would be recorded and tracked in the same system.

We decided to accomplish this by creating custom templates in JIRA [2] to track all of our digitization requests. Stanford Libraries had already been using JIRA as a tool for software development and bug tracking and the software was easily customizable so we could use it to record, assign, track all of our digitization requests.

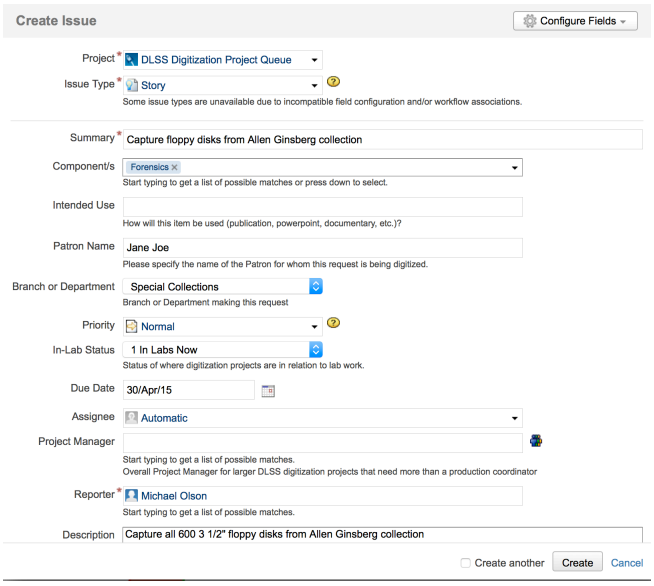
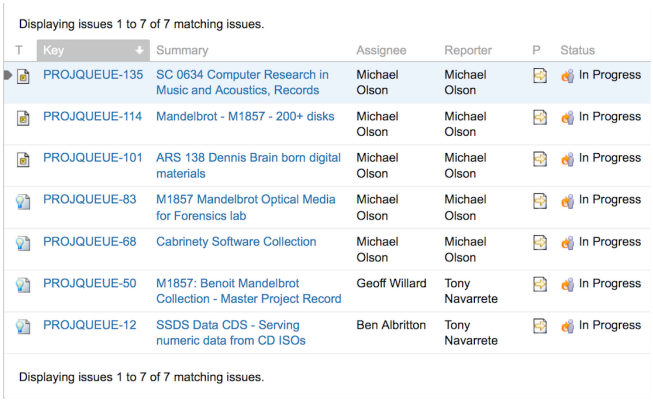



***Figure 1.*** *Sample JIRA Digitization Request Form*

There are three types of digitization request. The first is a "patron digitization request" that is submitted by public service librarians

on behalf of a researcher or student. An example of this kind of request is when a faculty member requests the digitization of map or a manuscript for their research and or publication. These requests are defined as requiring less than twenty hours of lab labor. The second type of request is called a "one-off on demand request". This type of request is initiated by subject librarians and is typically used for providing digitized materials for teaching. Similar to the "patron request" the "one-off internal requests" are typically small and require less than twenty hours of labor in our labs. The final type of request is called a "project request". Requests to digitize large volumes of library materials or recover data from obsolete computer media are initiated by subject librarians and as part of this process we ask the requesters to prioritize their project requests. This prioritization allows us to allocate limited lab throughput towards content that is identified as being the highest priority to preserve.

The use of JIRA for tracking our lab work has greatly enhanced the transparency of what is in our labs. Customizable reports are easy to generate and notifications can be used to inform clients of status updates.



Displaying issues 1 to 7 of 7 matching issues.

| T | Key | Summary | Assignee | Reporter | P | Status |
|---|---|---|---|---|---|---|
| | PROJQUEUE-135 | SC 0634 Computer Research in Music and Acoustics, Records | Michael Olson | Michael Olson | | In Progress |
| | PROJQUEUE-114 | Mandelbrot - M1857 - 200+ disks | Michael Olson | Michael Olson | | In Progress |
| | PROJQUEUE-101 | ARS 138 Dennis Brain born digital materials | Michael Olson | Michael Olson | | In Progress |
| | PROJQUEUE-83 | M1857 Mandelbrot Optical Media for Forensics lab | Michael Olson | Michael Olson | | In Progress |
| | PROJQUEUE-68 | Cabrinety Software Collection | Michael Olson | Michael Olson | | In Progress |
| | PROJQUEUE-50 | M1857: Benoit Mandelbrot Collection - Master Project Record | Geoff Willard | Tony Navarrete | | In Progress |
| | PROJQUEUE-12 | SSDS Data CDS - Serving numeric data from CD ISOs | Ben Albritton | Tony Navarrete | | In Progress |

Displaying issues 1 to 7 of 7 matching issues.

***Figure 2.*** *JIRA report of projects that are in process*

## Budget

Developing a budget is necessary to track and forecast lab expenses for past and future fiscal years and shows our return on investment. Beginning in fiscal year 2014 we began to track our anticipated and actual expenses. The expense categories we now track include workstations, other computer hardware components, consumable commodities such as imaging targets and general lab supplies, software, contracts and maintenance agreements for both hardware and software, hourly staffing, and travel. Our digital archivist is base funded so this salary expenditure is not included in our current lab budget. We also maintain an equipment replacement schedule that allows us to forecast when we will need to replace our workstations.

| **FY 2015 Roll-up (updated 03/02/2015)** | | |
|---|---|---|
| | **Projected** | **Actual Cost** |
| Equipment Labs | $4,720.00 | $1,443.99 |
| Equipment (Computing) | $2,687.10 | $1,160.22 |
| Consumables | $4,500.98 | $915.72 |
| Contracts & Maintenance | $9,958.47 | $844.37 |
| Software | $108.00 | $372.76 |
| One-time (Misc.) | | |
| Travel | $2,609.44 | $2,609.44 |
| Hourly Support Staff | $15,000.00 | $6,628.00 |
| | | |
| | **$39,583.99** | **$13,974.50** |
| | | |
| **FY 2014 Roll-up** | | |
| | **Projected** | **Actual Cost** |
| Equipment Labs | | $929.93 |
| Equipment (Computing) | | |
| Consumables | $1,000.00 | $488.40 |
| Contracts & Maintenance | $3,345.37 | $3,336.37 |
| Software | | |
| One-Time (Misc.) | | |
| Travel | | |
| Hourly Support Staff | $15,000.00 | $3,899.00 |
| | | |
| | **$19,345.37** | **$8,653.70** |

***Figure 3****. Born-Digital / Forensics Lab Budget Forecast.*

One very important budget line that has had a direct impact on our capacity has been allocating significant dollars to hire hourly labor. Hiring a lab assistant for nineteen hours a week consistently for fiscal year 2015 has lead to a significant increase in our lab throughput even though at the time of writing we are only half way through the current year. This has freed up our digital archivist to spend more time training project archivists in recovery of obsolete computer media they find in their collections. It has also allowed us to focus our digital archivist on more complex tasks such as recovering data from hard drives and working with prospective donors of born-digital content.

## Standardization of Equipment and Configurations

Currently our digital archivist is responsible for maintaining all of our lab equipment: eight workstations primarily devoted to disk imaging or capturing born-digital content, two FREDs for processing collections using forensic software tools, and a portable laptop for doing remote acquisitions. We are attempting to reduce the time our digital archivist spends maintaining this equipment. As our older equipment comes due for replacement we are no longer building our own custom workstations in favor of purchasing commodity workstations. We still purchase specialized equipment like FREDs and hardware write blockers but maintaining custom built solutions is no longer necessary. There are now multiple types of interfaces available for connecting modern workstations with older legacy drives allowing us to gradually replace our custom built capture stations with floppy disk controller chips with commodity equipment.

Creating disk images of our workstations is another way we are trying to reduce the time our digital archivist spends maintaining our lab equipment. Occasionally we are required to do complete restores of our lab workstations. When this is done manually it can be very time consuming and is not the most effective use our digital archivist's time. One of our goals for this year is to implement a plan for saving disk images of all of our workstations allowing us to more easily restore machines to working states with less hands on labor.

## Documenting Lab Procedures

The hiring of hourly lab assistants has greatly improved our efforts to document lab procedures. Utilizing an hourly or student labor force necessitates the creation of documentation as hourly staff are by their very nature temporary. Functional documentation makes it easier for us to train new hires and also reduces the risks associated with retaining too much knowledge in the mind of a single staff member.

## Born- Digital / Forensics Labs Statistics

We have gradually improved how we track born-digital materials that are preserved in our labs. Improved statistics have allowed us to identify process improvements in our workflows and show the cost and value of preserving born-digital collection materials.

As part of our statistics we document the following data points for all materials that are preserved in our labs:

- Collection call number
- Media type
- Quantity of media
- Failure rate by media type
- Total number of files (when a file system can be identified)
- Size (GB) of preserved content
- Operator responsible for capture
- Operator responsible for quality control
- Date that the capture of collection is complete
- Number of hours required for capture and quality control.

We are continuing to discover new data points that we would like to capture as part of our statistics. For example, we are preserving raw KryoFlux stream files when other forms of preservation fail. These raw stream files represent the magnetic flux read from the magnetic media and have no identifiable file system. We are preserving these raw stream files but have yet to determine how we should account for this in our lab statistics.

Stanford's yearly statistics follow our fiscal year beginning September 1st and ending August 31st. Although our current fiscal year is just over half complete the increased use of hourly labor has lead to a significant increase in the number of collections and the number of media preserved. We have focused the hourly lab assistant on collections that contain large quantities of floppy disks and or optical media. The skills required to capture these media types are easily transferred to hourly lab assistants and or students. Our digital archivist is responsible for supervising our hourly lab assistants and provides expertise with problematic media when required. This has freed up our digital archivist to focus on more complicated capture of hard drives and whole computers.

**Born-Digital/Forensics Labs Fiscal Year Statistics**

| **FY2015 Roll-up (updated 03/02/15)** | |
|---|---|
| | |
| No. of Collections | 15 |
| No. of Media | 2,096 |
| Files | 8,685,204 |
| Size (GB) | 10,726 |
| Staff Hours (casual) | 171 |
| | |
| **FY2015 Roll-up** | |
| | |
| No. of Collections | 6 |
| No. of Media | 455 |
| Files | 6,870,515 |
| Size (GB) | 9,967 |
| Staff Hours (casual) | 105 |

***Figure 4***. *Fiscal Year Lab Statistics*

Statistics from our Born-Digital / Forensics labs should be analyzed with caution. A single acquisition of a large collection can easily skew our statistics particularly when viewed over only a couple of year's data. Nevertheless, we anticipate that maintaining lab statistics over many years will allow us to better anticipate future long-term preservation storage needs. We also anticipate that we will be able to better provide our librarians with more accurate estimates on the time and dollar resources required to preserve born-digital collections.

## Conclusions

In the past two years Stanford has implemented a number of significant changes that have allowed us to turn our Born-Digital / Forensics Labs into more of a production facility with greater capacity to preserve the increasing quantity of acquired digital collections.

Implementing a clear and defined process for requesting lab services, prioritizing collections for capture and having a documented process or lab workflow sets the stage for increased productivity. Defined and documented lab operations are repeatable and therefore scalable. It has also had the benefit of increasing the transparency of the materials that our flowing through our labs and how we are serving our librarians and research patrons.

A documented lab budget has greatly increased our ability to serve our clients and demonstrate the value of our services. It has also allowed us to better estimate the costs of preserving born-digital collections and to expose these costs to the librarians that are responsible for their acquisition.

A shift to standardize our lab workstations and move away from custom-built solutions is anticipated to decrease the overhead

in managing a suite of more than ten workstations. Our goal this year is to create disk images for all of our workstations and reduce the time consuming work that is currently the responsibility of our digital archivist.

Our increased use of hourly labor specifically focused on capturing floppy disk and optical media has necessitated the creation of functional documentation that describes the work we do in our labs and makes it easier to train new or additional staff. This use of hourly labor has also allowed us to increase our lab throughput and provide preservation services to a greater volume of born-digital collection materials that are at risk of loss. There still remains much we can do to improve the functionality and throughput of Stanford's Born-Digital / Forensics Labs but continuing to better define our service, track our work, and look for ways to improve our productive throughput have put us on a path to better meet the needs of our librarians and research patrons.

## References


[1] KyroFlux. http://www.kryoflux.com/

[2] JIRA Issue and Project Tracking Software. https://www.atlassian.com/software/jira


## Author Biography

*Michael Glenn Olson is the Service Manager for Stanford's Born-Digital / Forensics Labs. In this role he is responsible for steering the technical development of the labs and formulating the service model for preserving born-digital data. Michael has a BA in Medieval Studies from the University of British Columbia (1997) and his M.Phil in History and Computing from the University of Glasgow (2000).*