Digital Curation: Two Paths
Stanford University Libraries

Society of California Archivists
April 28, 2012
Ventura, California

Glynn Edwards, Librarian for Digital Archives & Manuscripts Programs
Department of Special Collections & University Archives
This morning I will present a brief overview of the (born) Digital Archives Program at the SUL.
Mine is relatively general and will (hopefully) provide the context needed for Peter’s more detailed and technical focus. As you would imagine a lot has changed since the completion of the AIMS Project in the winter.
One of our tasks this winter and spring has been to begin crafting our Mission Statement within SULAIR and at SU. Part of this is contingent on funding to build out the lab (equipment and staff).

The DAP is a hybrid program within the libraries. In some ways it is LIKE other data groups – like Branner Library which assists with GIS data or SSDS which assists with data sets. It is also like the digitization labs at SUL – the SMPL (Stanford Media Preservation Lab) and Map Lab, etc. BUT it is also an acquisition and processing unit – similar to our Manuscripts and Archival units – working in conjunction with other staff and departments within Special Collections, DLSS, etc.
This slide illustrates an early DRAFT for our “service level agreement” – describing to our community what the DAP / forensic lab does.

But obviously we need to better understand and consider the following:
Understanding the community we serve
Describing the formats we capture from, the hardware, software and tools available
What type of training and/or workshops might be available
When we offer one-on-one assistance and what that encompasses

Questions: Will we attempt any cost recover? Will we capture data for personal use on campus?
What services can we offer with existing staff model?
Now – onto our current high-level workflows – of which we have several!

High Level Workflows: based on different record formats
This is our basic FTK workflow for processing – primarily text- or mixed-collections. It was developed during our 2 years on the AIMS Project working with (admittedly) small born-digital materials from larger paper collections.

In our work since AIMS though we have found that FTK does not work best for ALL file formats – and other tools were found and workflows were developed.
For our digital photography collections (or large subsets) – when reformatting labs or the photographer has stored metadata in the IPTC header embedded in the digital object;

We found that that metadata was not read by FTK. So, while FTK is part of this process it is not the MAIN processing tool for this format. Instead, after testing, we selected PhotoMechanic software.
Here is an example of PhotoMechanic – these are screen shots from the Douglas Menuez collection.

The software allows us to create and use templates and to add bulk metadata across selected images.

The files and metadata are scripted out to MODS for ingestion into the SDR (Stanford Digital Repository).

This is the format that we are currently delivering online – because it is so similar to digitized photographs.
Another format that gave us difficulties in FTK were Email collections. If one email was flagged as restricted, all the emails in the “box” were flagged as restricted. There was no way to pull out one email or one thread.

Conveniently, Peter discovered a software in development at SU’s MobiSocial Lab in Computer Sciences – called MUSE. It was developed by Sudheendra Hangal whom we have worked closely with this past 6 months or more. I will show you a few screen shots of MUSE later.
But first, I would like to talk a little about our digital repository. Output from our various processes needs to be ingested in the SDR for long-term preservation and for delivery.
Here is a snapshot overview of the SDR triumvirate of Management, Access, and Preservation.
Delivery is via persistent URL’s (PURLs) which resolve to “whatever” delivery environment is used for that “format”

Delivery for born-digital content in the short term is more complicated. You will note the orange oval – which sits outside of the SDR infrastructure. In it I have put the three core processing tools there as they are not yet included in the suite of tools that map data to DOR/SDR.
Digital photographs (PhotoMechanic data) is the only format (so far) that has been ingested into the SDR.

This is a screen shot of our Management / Administrative interface – ARGO – for the digital photos in Fuller.
If we click on one image, you can see the Datastreams that are tied to this object; the unique IDs including a “local” one we use as a mnemonic for the physical location of the original.

ARGO object page also gives us a link to the PURL landing page for the object.
Discovery and access to the digital photographs is via our online catalog (SearchWorks).

The collection-level record includes a link to the set of digital objects; but the objects themselves come up as well if you search for that collection.
While DLSS (Digital Libraries Systems & Services Dept. in SUL) is working on creating a data model for ingesting FTK output – slated to be completed this summer ...

What this means currently is that access to files processed on the AIMS Project can only be delivered via FTK by appointment with our resident Digital Archivist – Peter Chan.
This screenshot is for a prototype – called Hypatia – developed at the tail end of the AIMS Project last fall.

Collection-level records were ingested along with a set of photographs of the physical media – but not the files themselves.
I’d like to switch now to MUSE again – and show you a few quick screen shots.

Besides being used and tested for Processing Email Archives, MUSE will be tested as an appraisal tool, a pre-accessioning tool, and for delivery via the reading room in Special Collections.
One of the underlying structures for MUSE is the built-in Lexicon – which is using natural language processing to group “sentiments”

In working with Sudheendra, we requested the ability to build our own Lexicons or to edit the existing one. That capability was added and MUSE will now accept multiple lexicons.

This is one Peter created for one of our collections
This final view of MUSE – shows two visualizations based on the contents/data.

The first is the Sentiment analysis based on full-text analysis of similarly grouped terms – for family, emotions, etc.
If you click on any of the colors (here presented over time) you will bring up the actual emails associated with that “emotion” and can scroll through them.

The second is Correspondents over time – which pulls data from the header information.
After AIMS was over, and with Peter now base-funded, we began setting ATTAINABLE monthly targets – or goals as well as longer term projects and priorities.

Some examples of monthly targets are:
- Capturing data from hard drives, computers and virtual transfers
- Assist with identification of computer artifacts – as in the Jerry Manock papers

Longer term goals are:
- Test and revise workflows
- Continue to seek out and test new tools, etc.
- Build relationships with faculty, curators, student groups, etc.
Current activities & issues

1. Pre-accessioning activities
2. Going into production: Stop Aids Project
3. Training lab & project staff
4. Capturing Rebecca Solnit’s 3 laptops (pix)
5. Post-accessioning reviews with curators

Our work often gets sidetracked by issues or problems encountered while trying to capture hard drives. Here are three laptops from the Rebecca Solnit collection – one, an old MAC laptop, requires a cable/connector we did not receive nor have.

Peter has to be very creative in trying to resolve these issues – often going far from Stanford for help.
Another project came about when one of our curators was taking in a congressional collection.

We created a survey / questionnaire for staffers to fill in to assist with understanding the scope, files, extent, etc. of the digital material. I posted it on the congressional papers roundtable listserv and rec’d interest but most did not hold out much hope of getting responses.

In this example – we rec’d a very detailed 5 page response – which has been a wonderful resource for us as we proceed with his digital component.
After AIMS we rec'd a grant to process the Stop AIDS Project records and digital content.

SAP was the production “test” of the workflows we’d developed earlier.

At this point the data has been captured and awaits processing.

Some interesting stats though:
Data loss from CDs was very high! These were back-ups primarily but we were only able to capture about 4% of hundreds. 3.” floppies had a success rate of 60% and zip disks 96% capture.

Dataset is nearly 500K files and over 850 GB – much larger than the “AIMS” collections.
As we needed to process the digital content as part of the grant, Peter had to quickly pull together training sessions for staff.

He developed FOUR workshops – attended by SAP staff and other Special Collections and Preserving Virtual Worlds staff

Our goal is to consolidate a lot of this material on our website over the summer.
The processing of the SAP records this summer will test our initial workflows greatly.

The size of the digital collection far surpasses the AIMS collections in complexity and number of files.

Descriptive Metadata: standards and discovery

- How is it entered? Machine? Person?
- When is it entered – pre-, during or post-processing?
- Where is it discovered? EAD? Web? Reading room?
- Standards – LC subject headings? Keywords? Tags?
- Entities – how do we reconcile versions of names, etc.?
- EXTENT will be a huge factor!
What does the future hold?

• Delivery - not yet where we’d like it
• Determine priorities and staffing for:
  – Capture from media (backlog & incoming)
  – Processing (growing backlog of accn’d material)
• Expansion of Lab capacity and capability

The ultimate goal of delivery would be to discover across collections and formats. That said we are far from searching across digitized and born-digital metadata and full-text at this point.

We also need to determine how we will: capture data from our growing backlog of physical media; process the data (staff and workflows)

It all hinges on how our lab is staffed and our capability to expand capacity and tools as needed.
The team involved on the AIMS Project form the core of the DAP – Peter, Michael, me. To that we have added the expertise of Henry Lowood and Daniel Hartwig.