Introduce myself (day job). Before we get started, I want to just give you a quick intro into how I started working with Born Digital content

- Start working on the preservation of born digital content purely through necessity.
  - In early 2013, the Stanford Archive of Recorded Sound, where I was Sound Archives Librarian at the time before moving to Yale where I work now, started to receive a number of born digital audio and video collections. This was initiated by one of the annual deposits made to the Archive from the Monterey Jazz Festival.
  
  - Stanford holds the Festival’s complete recording archive and annually takes in the festival’s most recent set of live recordings, both audio and video. Until 2013, materials had come to us mostly on CD, DVDs, DV Tapes etc in multiple boxes. However, suddenly we were now starting to receive just a single box containing two hard drives, one with audio files, the other with video files. So I started to wondered what we could do with these.

- Truth be told, we had in fact received a few other born
digital collection items previous to this, often mixed in with other collections of analog materials. However, like many other archives, we had treated these digital collections like any other archival collection material...we had processed it, stuck the hard drive in a box, and planned on dealing with access in the future!

So I decided now was the time to start tackling the preservation of digital collections at the Archive of Recorded Sound. I will fully admit that when I started working these collections I had pretty much no idea what I was doing! And while I certainly now don't claim to know everything about working with born digital content I am now a lot further forward in my knowledge mainly thanks to taking that initial plunge into tackling these collections and not leaving them.

So if there is one message I want to ring out loud and true today and it is “don't be scared” “dive in” and start working on your born digital collections now. Putting hard drives in an acid free box as you would with papers is not sufficient for born digital content. Stabilizing the physical media is not the same as stabilizing the content.

I know one thing I was worried about when I first started was “what if I somehow delete all the content on the drive!” Please to say that I have not done this so far! I am going to give you a few pointers today that will ensure that you don't do this!

There are two things that helped me enormously to get up to speed with how to work with born digital collections.

The first was the Society of American Archivists’ Digital Archives Specialist certification program, which I completed this time last year in fact when I flew to Austin for 24 hours from Orlando in the midst of the MLA meeting last year to sit the final exam! That was a crazy week. This program, which is a multi module program that has to be completed within a two year period, is the benchmark for professional development in relation to digital archiving in the US at least. I found this course to be very rewarding and enabled me to connect with the current best practices in the field. Even if you don't take the whole program, maybe doing one or two of the courses would be helpful. You could even ask around at your institutions to see if there is interest in holding one of the courses at your
○ library as a group. SAA are very open to doing this. Yale has done this for 2-3 courses now.

○ Second, I have been fortunate to work at two institutions, Stanford and Yale, with fairly well established programs related to born digital content, and noted experts in the field on staff. This has allowed me to learn on the job by being part of the process of working with born digital collections alongside people who are much more experienced or knowledgeable in the field that I am.

○ Maybe mention NDSR too There is also the National Digital Stewardship Residency (NDSR), run through the LOC, which is a residency program for recent grads where they learn alongside pros in the field at various places across the country.
Enough **CORE KNOWLEDGE** to start using **TOOLS** and create **WORKFLOWS** to process born-digital content

- My objective today is to give you enough **CORE KNOWLEDGE** to start using **TOOLS** and create **WORKFLOWS** to process born digital content

- Will touch on some theory, but basically want to give you some **practical advice and hands on experience** to help you process that first collection if you have not done so already and to make the idea of working with born digital archives a little less scary!

- Won't be touching on descriptive metadata (e.g. description in a finding aid) in much detail. However, we will discuss the need to capture a **fairly significant amount of metadata and documentation during the process of accessioning and basic processing** of born digital content that will significantly inform your later description work for that collection.

- Another area I am not going to talk about much is the **pre donation / pre transfer appraisal of born digital materials**. However, to offer a few words of advice about this stage
  - Get you digital / technical staff involved in this process **ASAP**
  - A small hard drive does not equal a small collection!
You need to consider

- **Storage infrastructure for storing the files.** Where are you going to store digital content that you process? Digital storage is ever cheaper, but there is a $ cost for storage you need to consider.

- You need **policies** in place that outline what **file formats you can store and make accessible**, as well as what you will do with any **delete files or track changes in documents that a donor might give you, often inadvertently.** Will explain more later.

- **Focus on materials given to the archive on physical media,** however will touch on materials from a institutional server and for materials on a donors machine and how to properly send these.
One caveat I will mention here is that there is an argument for applying the same workflows we will discuss today to made-digital content that comes into the archive for which you don’t hold the analog source but have simply acquired the digital surrogates. It is possible that the original sources could be lost etc and therefore the made digital copies are the only copies that survive, which could therefore warrant the same level of treatment as unique born digital collections.

CORE KNOWLEDGE: What is born-digital content?

“Digital materials which are not intended to have an analogue equivalent, either as the originating source or as a result of conversion to analogue form. This term has been used...to differentiate them from 1) digital materials which have been created as a result of converting analogue originals; and 2) digital materials, which may have originated from a digital source but have been printed to paper, e.g. some electronic records.”

**CORE KNOWLEDGE:** What is digital preservation?

A “series of managed activities necessary to ensure continued access to digital materials for as long as necessary.”


- **Read quote**
  
  This means that when we speak about digital preservation, we are talking about not only securely capturing and then storing digital content initially, but then monitoring the content consistently to ensure it is accessible in the future and to guard against file corruption and obsolesce.

- **Adding files to a digital archive** or preservation system is the digital equivalent to arranging physical special collections materials in acid free housing and storing them in a climate controlled and actively monitored environment, or long term preservation.

- **Digital preservation systems** store multiple copies of files, on multiple different types of storage media (primarily hard disc and data tape), in multiple geographic locations.
  - The more exact copies you have in different places, the better chance you have of recovering a copy if say some storage device fails or a files gets corrupted or a data center goes down etc

- Options for preservation systems are either **vendor products like Preservica**, which we use at Yale, or **local installs of programs like**
Archivematica, which is an open source suite of tools covering ingest to access that also integrates with things like ASpace etc. Or you might have a totally home grown system like Stanford.
Checking the fixity of digital content is one of the core pillars of digital preservation. In short, what you are looking to ensure here is the integrity of the objects by proving that the files you store in the archive, say on a backed up server or digital archive / preservation system, are exactly the same as the files you received from the donor and have not been altered or corrupted since they came into your custody.

Additionally, I will mention that in general born digital archiving is no different to paper archives. You still want to document and ensure the provenance, original order and chain of custody.
In practical terms, Fixity is ensured by verifying a file’s checksum. A checksum is “a unique numerical signature derived from a file.”

This is a stamp in essence baked into the file.

If any change is made to the file, no matter how small, the file’s checksum will change.

This includes if the file becomes corrupted. So if you want to ensure that one copy of a file is the exact same as another, as contents are transferred between storage media throughout the stages involved in processing born digital content, i.e. coping materials off of an external drive and onto a server etc, then verifying that the checksum of one file is the same as the other is how to we ensure that those two files are exact copies of each other.

Digital preservation systems will regularly run hash checks to establish fixity of content.

Show how MD5 values of a file change when you alter a file.
One thing we will talk about as part of the processing workflow we will discuss is the need to **capture and document these checksums as part of our work**.
Another thing we need to ensure that we don’t alter when processing born digital content are the MAC dates of files as this is provenance information. These are the modified, accessed and created dates of files. We also want to capture these dates for each file as part of our processing work.

Also, this is the date info about the objects you will need when describing the collection!
• **Read Quote**

• **Think about this, for those of you from the cassette or floppy disc era, this is the digital equivalent to knocking out the tab on a hard drive, to make sure nothing can be written to the drive. It means the flow of data communication is just a single direction from the drive to the machine. The machine cannot push any information or change anything on the drive, no matter how small.**

• **One mistake people make when receiving a digital collection is to just plug the drive in and start looking around.**

• **Even the act of just plugging the drive in, without a write blocker, could alter or even corrupt files in extreme circumstances.**

• **Do not start trying to access or in anyway survey a collections content on a drive until you have processed and stabilized that content, which we will talk about.**

• **Hardware writeblockers are best practice but are expensive and can slow up the transfer process.**
• Good software blockers are also pricey. So this might be one area you need to invest some $$$.
Best free one is the one that comes with bitcurator, which we will use today.
You will hear the acronym PII a lot in the literature about born digital processing. This is Personal Identifying Information. This includes SSN, Bank accounts, email addresses, phone numbers etc.

Locating this data is easier with born digital as you can use tools to scan for this information across a body of files.

If you find this info, you may need to redact it before making the files available. If not critical resource, maybe consider deleting it.
• **Dont get too scared by this - final bit of knowledge**

• Hexadecimal, often just referred to as HEX, is an alphanumeric system commonly used to represent binary code, i.e. the ones and zeros that make up a digital file.

• **It is a human readable form of binary**

• **Digital preservation is focused on preserving files at the bit level, the root level, not just at the level that a file system might render a file.**

• Hex allows for common patterns of 1s and zeros to be easily identified. A good example is the opening 25 50 44 46 pattern on row 0 here. This is often known as the magic number which frequently allows you to identify the binary pattern that represents the file format. In this case we have 25 50 44 46 which represents the pattern associated with a PDF file.

• **Dont want to spend too much time on this, but it is useful to have a base understanding of this as the tools we will use today interact with files at this level and use this data to identify file formats, PII etc etc.**
• Any questions before I move on. **More of this will make sense when we starting working** with tools and try to process content.

• Two of the major tools used for processing born-digital content are BitCurator and FTK, produced by a company called AccessData.

• FTK is used by many institutions and is the gold standard in essence for processing born digital materials.
  ○ While used by the Library and Archives community, it was designed and more targeted by the law enforcement community, and is the main tool used by most digital forensics departments.
  ○ However a **single perpetual license is $4000** with another $1200 a year for support so it is pretty expensive.
  ○ AccessData do however produce a free stand alone application for creating, mounting and inspecting disc images however, called FTK imager. It's not all that powerful however and is only available for windows.

• The other tool, and the one we will be using today, is called **BitCurator**.
This is a complete suite of open source tools produced by the **BitCurator Consortium**.
- **Designed by and for the library and archives community** it was developed as a collaborative effort between the School of Information and Library Science at UNC Chapel Hill and the Maryland Institute for Technology in the Humanities (MITH).
- They created some of the tools, others are open source tools produced by third parties that have been packaged together.
- **Offered as an alternative to the expensive FTK suite.** However, arguably not as powerful and FTK includes more powerful search capabilities and better GUI interfaces and analysis tools.
- **Bitcurator Consortium is a group of member institutions** so it is supported by a fairly large community of institutions. The community or users is fairly large and active too.
DOCUMENTATION

- Documenting your actions and decisions at every stage is critical.
- This will help record the preservation activities on the content and also help inform future workflows.
- It will notably provide evidence of the steps you took to ensure the provenance and authenticity of materials throughout the processing period.
- Link documentation to accession record if possible
- Keep documentation in project folder with content etc

ACCESSION RECORD

- You still need to create an accession record at the beginning as you would for any archival collection. However, it will be fairly minimal for now and you will then expand upon it after the processing is done.
- You would just include a title, extent (hard drive), info on the donor,
and a scope and content notes of what you think is on the drive based on donor supplied into. You again will update this all at the end of the process.

Establish an accession number - mlatechcamp2018

PROJECT WORKSTATION

You need a dedicated workstation for processing digital content. This should be:

- Regularly scanned for viruses. Making sure this machine is virus free is critical
- Keep all software up to date
- Have a good amount of memory and CPU power
- Keep non-networked until a network connection is needed, i.e. for virus definition updates or file transfers. Reduced the vulnerability of you machine as it is closed off. Also prevents unauthorized network access, especially important for restricted collections
- Make sure the machine has a good number of connection ports, USB, Firewire etc and also that you keep a selection of cables to connect drives
- This workstation should be thought of as a quarantine zone (which we might use for paper collections that have not be checked for mold etc)
- Also forensic processes can take time, so the machine will be occupied and could not be used for other tasks

So, we dont have this ideal setup today of course, although hopefully most people's machines are updated and virus free. We are connected to a network too, but that's not going to stop us!

Before you do so, close all other running apps on your machine except Slack. Bitcurator needs a lot of RAM memory and CPU, so they more you have available the better

So if you have not done so already, go ahead and boot up the Bitcurator VM in Virtualbox. If you have any problems with this loading, just put up your hand and one of the ETSC members will come and try to help you out.

Once you have opened up Bitcurator, the first thing we are going to do is create a folder for this processing project. This folder will contain
all of the transferred content and documentation
  ○ Right click on desktop
  ○ Create top directory - mlatechcamp2018
    ■ I tend to name the top folder for this directory using the **accession number** assigned when creating the accession record. However, you may wish to use **some other naming convention**. However, make this convention consistent when processing collection and **make sure the names are unique and obviously link the material** with a collection
  ○ Under this top folder, go ahead and create three subfolders called, master, working copies and documentation.

**CONNECTING MEDIA**

- So we are now ready to connect our media! But before we do, what do we need to do?? **Writeblockers**

- **Our budget for the #techcamp did not extent to buying 60 or so** hardware write blockers. However, there is a software blocker built into **Bitcurator**
  - **Demonstrate this and the two states**

- So make sure the **green light is on and connect your USB stick**

- **Still dont open any content**

**TRANSFER MEDIA**

- **OK so we are now ready to start transferring content off the USB drive**

- There are **two basic ways** to transfer content
  - Disc imaging
  - Copying files using a standard called **bagger**

- **We are going to focus on disc imaging today, but I will talk about bagger if there is time later.**

- Disc imaging makes an **exact bit level copy of a physical drive**
  - Disc imaging ensure that **all original metadata along with deleted files and info on the file system of the drive itself**
  - Disc imaging in essence creates an **exact snapshot of a drive**
and its content, thereby stabilizing the drive by retaining an exact copy of it to move to a different storage environment.

Two types of images
- Forensic images which capture everything including delete files and unallocated space
- Logical images that omit unallocated space and deleted files, just capturing in essence the active content

Deleted files and track changes etc
- Explain basic file system method of dealing with deleted files
- Need a policy in place that discusses what you will do with deleted files or if you will even attempt to recover these. Keeping large forensic images will take up a lot of storage space, so you need to think through what you approach to this will be

Disc imaging in Bitcurator is done using a tool called Guymager
- This is in the imaging and recovery folder on the desktop
- Open and right click on the USB drive
- Many types of disc image, but today we will use today is dd. Expert witness does allow for you to add some additional metadata. But it is a proprietary format. So we prefer to use non-prop such as raw linux image.
- Select a location for your image home > bcadmin > Desktop > mlatechcamp > masters
- Create a name for your image, again be consistent here so maybe accession number_and disk number. It it is good practice to add a physical label to the drive with this on it, so you can match the drive and the disc image.
- Calculate hashtag and I would suggest verify after too.
  - This uses checksum checks to verify that all the data in the disc images matches the data on the drive, again ensuring it is therefore an exact copy
- Start.

Will hopefully be quick for our small drives but large drives can take a very long time.
Once complete, right click and unmount and then eject.

Thats it, you have a disc image

Look at image file and info file data

- Link to copy of disc image for anyone who has issues with imaging
  https://yale.box.com/shared/static/xd4zbcrzjadp55nwoyymbocr1bjsoqul.e01

REPORTING

Creating
- So we have now created a snapshot of the content from the drive, which is the first step in stabilizing it

- However, we still don't know anything about the content really. Before we start explore (copies) of the files we need to
  - Establish what file formats are in the content
  - Check to see if there is any PII in any of the files that we need note to redact or check with the donor
  - Document all of the checksums for the files as a way of ensuring authenticity going forward, for fixity checking coping etc
  - MAC dates
  - Directory organization

- This information will be added to our documentation and some of it will be included in our accession record and later in our archival description

- This can be done via the Bitcurator Reporting Tool

  - Go to Forensics and Reporting > Bitcurator Reporting Tool

  - Select the Run All tab

  - The reporting tool is made up of three major components
    - FiWalk - this is a forensics analysis tool which scans a disc image and creates a structured xml file detailing
      - Volumes
      - Directories
      - file structure of the content
      - a list of the files
• file sizes
• checksums for each file
• file formats
• and MAC dates
• So this is the bulk to the data we want to know about the files.

■ **Bulk Extractor** - another forensics tool, this one however is more **concern with content than file system info**. We will be using bulk extractor to look for PII.

■ **Reports** - FIWalk and Bulk Extractor create useful data that is useful to a machine to parse, **but not easily human readable**. The last part of the reporting tool therefore creates **PDF and excel copies of the data which are easier to read and will be included in our documentation**

  ○ First thing we need to do it **create bulk extractor data**

  ■ **Create a folder in your documentation** called Bulk Extractor Output

  ■ **Click on the BE Viewer in reporting tool**

  ■ **Go to Tools > run bulk extractor**

  ○ **Select your image file**

  ○ Then **select your output** for the extracted data.

  ○ **Many types of scanner**, but we are going to just run the default and then I will show which are most useful to us.

  ○ Then hit **run** and let is do its thing

  ○ Once you see **done, click close**

• **Go back to the reporting tool**

• **Select image**

• **Select Bulk Extractor Folder** in next field
● **Output Directory** - point to documentation folder and give your folder a name reports

● Click Run

● This can take a few minutes. **Will take hours and maybe days for large images. Other reason to have a dedicated machine.**

Evaluating

● **So lets take a look at the reports** created to see what they include

  ○ The first one I want to focus on is the **FIWALK excel and XML file**. This is your main **provenance info and your documentation of authenticity**. You can also **use this to update you accession record**. This includes
    ■ **Full inventory of items, including any deleted files**
    ■ **Outlines directory structure**
    ■ **File formats** (Uses Droid tool developed by UK National Archives that uses the Pronom technical registry of file formats)
      ● This information will help you determine whether you will be able open and read the contents of digital files.
      ● Also add to descriptive metadata to outline what might be needed to open the files
    ■ **Checksums**
    ■ **MAC dates**
      ● So these might look odd. This is due to the way I copied files onto these drives. I did this a few days ago. The tool I used retains the last modification date but not creation or accessed dates. But the last modification date is enough to give us the dates we need for our description work.
    ■ **Data size** (you will work to replace the extent in your accession record for MBs and just add the fact that the files were sent on a hard drive to a note field)

  ○ Go through FIWALK XML and excel

  ○ Look at PDF reports
- Look at PII excel reports
- Also includes PREMIS metadata which outlines the preservation activities performed on the files by FIWalk and Bulk Extractor

- During this process consider the sustainability of the file formats
  - As software changes over time, or is no longer supported, software dependent files can be at risk of becoming unusable
  - The Library of Congress’ National Digital Information Infrastructure and Preservation Program (NDIIPP) has resources on the sustainability of digital formats, including detailed format descriptions and other information that can help you decide if you should maybe create copies of certain files using lower-risk formats. I.e Apple Works to PDF perhaps.
  - This is a process known as normalization. It is worth noting that the more formats your organization chooses to maintain, the more effort will be needed over time to maintain them. So again, you need a policy really that outlines what formats you are willing to take.
  - An alternative to normalization, or if normalization is not possible, is to find copy of the original software and platform that ran the software and run an emulation of that platform. Discuss EAAS.

APPRAISAL

- So the next step is to start looking at the contents so we can try and figure out what all of this is. You will notice that until this point we have just been working on preserving the data, not really diving in to see what we have.

- One important step here is to work with copies of the content and not the preservation masters. This will ensure you can do whatever you want with the files without affecting the original content and metadata.

- So to do this, we will open up Forensics and Reporting > Disc Image Access. Click Open Disc Image, and open your disc image file. So here
• you will now see the complete directory structure for the disc.
  ○ Files in red are deleted.
  ○ Files with dots in front of them are hidden files, not visible in your explorer.

• Go ahead and select all and export to your working files directory

• Navigate to these files and you can start working with them.

• Exported items are newly created copies so have new MAC dates
  ○ However, you original MAC dates are still stored in the FlwalkK data you had before when creating the disc image, along with your checksums.
  ○ If you export hidden files you will need to tell whatever explorer you are using to show hidden files or they won’t show.

• If you have the tools on you workstation to render the files then great. There are a number of tools built into bitcurator to access files too e.g VLC for media. But you might not, so consider quickview plus

• Other tools that exist in bitcurator include tools for:
  ○ Identifying and delete exact duplicates (using checksum)
  ○ Explore embedded metadata in image files (compression, color profile, dimensions, resolution PPI etc)
  ○ Read outlook pst files
  ○ Recover damaged files
  ○ Analysis if damaged discs

VIRUS SCAN
• In addition to regularly scanning your workstation, you will need to scan every batch of copied files to check for viruses. To avoid any alterations to the master files or metadata, run virus software only on a working copy generated from the master copy.

• When you find a virus, the next steps will vary depending on the nature of the content and the virus. One approach is to note the names and locations of the infections and, if possible, obtain a clean copy from the donor. Another approach is to create a “cleaned” copy using antivirus software (be sure to document this action). Keep infected files quarantined until you determine the best course of action.
- So before you move any files off your quarantined machine, you will want to **virus scan them**. If you are moving the whole disc image, you will need to scan copies of all of the files.

- Run through virus scan using Clam TK

**MOVE TO STORAGE**

- **Now one option here is just to keep the disc image as is.**
  - This would be the equivalent to **minimal processing** in digital land to my mind. You have:
    - stabilized the contents by taking it off removable media
    - have provenance and preservation metadata on the content
    - some basic information to provide a descriptive summary
    - Linked this Documentation in your accession record.

- This is the **equivalent of arranging and minimally describing a paper collection.**
  - You have **not weeded, or described further** than that.
  - However, large disc images take up storage space, and mounting a large image to make it accessible to a patron, especially over a network if your image is on a network server, is slow. You would need to **pull it down off the network first.**

- So one option might be to mount the image, copy the files you want to keep, and discard the image. Use GRSYNC for this and use checksum option and preserve modification dates.

- So last step is to **move the content off of your workstation** and add it to a **secured backed up network server**, or to a **digital archive or preservation system**. Might be added to network server in the **interim** before it goes into your digital preservation system.

- **When you move files off of your dedicated workstation, be sure to checksum the content.** Use GRSYNC for this.

**TRANSFER CONTENT (BAGGER) - If time**

- As I mentioned earlier, another method of transferring content is bagger. This is very useful if you have a small amount of known content
- and you are not interested in deleted files etc.
- Put USB stick back in
- Double click to mount
- Open bagger
- Create new bag (no profile)
- Add files
  - Open media
  - Sdb1
  - Select mlatechcamp folder
- Allows you to add some basic metadata describing the materials
  - Contact name
  - Bagging date
- Click save bag as
- Locate to source to save
- “Holey Bag?” should be unchecked (a holey bag is a bag with an empty “data” directory)
- Save
- “Validate Bag” at this point to double check that all the files are present in the new bag and all the checksums are correct. It should be, but sometimes it feels good to be sure.
- Close and go and look at the bag
- See you get in terms of reports - checksums and filenames, so enough for fixity, but not much of file names etc. But you can maybe just look at the by hand.
- So that content and fixit data are now packaged together and that bag can be validated at any time to make sure the content file checksums match up with the manifest
- After this you can use GRSYNC to make a copy of the data folder for appraisal. Validate the original file afterwards.

- Can also use bagger to bag up content from a disc image if you are not going to keep the disc image.
Resources

Online

- Forensics Wiki http://forensicswiki.org
- BitCurator Wiki - https://wiki.bitcurator.net
- UC Guidelines for Born-Digital Archival Description https://github.com/uc-borndigital-ckg/uc-guidelines

Reading

- AIMS http://dcs.library.virginia.edu/files/2013/02/AIMS_final_text.pdf (chapters 2 and 3)
Questions

@jonmanton
jonathan.manton@yale.edu