Slide 1: Institutional Context:

1. To provide some background information on the Library, the University Library consists of several special collections and technical services units, and individual libraries. So even though we call it the University Library (singular), it is composed of not just one library – but a network of libraries.
2. The unit where I work is Preservation Services which is part of the Office of Digital Strategies.
    a. Preservation services is a centralized service area that collaborates with library collecting units to ensure persistent access to the intellectual contents of Library's collections of enduring value.
    b. Work closely across units with those responsible for the primary acquisition and stewardship of collections – archivists, curators, etc.
    c. Most often serves special collection units including:
        i. University Archives
        ii. Sousa Archives and Center for American Music
    d. I also work with other library service units such as Digitization Services and Library IT

Slide 2: Digital Preservation At Illinois:

1. One subunit service that I manage is the born-digital reformatting lab where primarily special collections content is migrated from fragile computer media such as floppy disks and older internal hard drives using digital forensics techniques.
2. At present, our digital preservation services only programmatically support bit-level preservation – we're recovering the data from aging and vulnerable media carriers and moving them to the managed and redundant network storage environment of our digital preservation repository, Medusa - a homegrown storage repository and associated collections management database.

3. With the contents stored in Medusa have access to the “bits on the disk” to continue refining and developing our programmatic digital preservation efforts.
4. The collection content recovered from legacy media often represents heterogenous group of files – ranging from unique manuscript documents of which this may represent the sole copy – to operating systems and other computer system related files – which can provide useful information to researchers and those curating the collections.

Slide 3 Software Dependency:

1. The files recovered from these obsolete carriers are often of the same era, or older, as the carriers themselves, thus subject to technological obsolescence, and dependence on legacy software.
2. These born-digital collections objects are often dependent upon obsolete software, or legacy versions of contemporary software titles where supported functionality and features often vary among versions of the same title.
3. Current versions of software titles indeed may not consider backward compatibility or only provides backward compatibility to limited set. Software producers are not in the market to have their software work indefinitely and on all versions. It isn’t a business model that they’re particularly interested in. Image one is a graphical representation of illustrating compatibility between files created in the Adobe Creative Cloud suite of tools.
4. Another example is the Pro Tools digital audio workstation software suite has at least three different session file formats. Image two illustrates the three-file format and notes the oldest version of the Pro Tools software that can open those sessions.

Slide 4: Initial Engagement with Emulation and Software Preservation:

- Prior to 2018, I understood the importance of software preservation and emulation to digital preservation. We undertook few forays into researching and applying emulation as an access and preservation

strategy. However, we too encountered common roadblocks of lack of resources and scalable solutions.

- Although accepted as a digital preservation strategy, at present implementation and use is often limited to research projects or to institutions that have a great deal of resources dedicated to digital preservation
- Widespread and scalable use limited as there are steep resource barriers to entry
- Emulation often require significant technological knowledge and administrative resources to research, develop and implement solutions. Efforts beyond project-based or research efforts have not proven scalable.
- As software themselves, emulators are subject to the same technological obsolescence as other software titles. Thus, they will decay and lose functionality over time to keep them functional in contemporary computing environments.
- Efforts centered around specialized implementations are often too resource intensive and not scalable enough for practical implementation
- Despite the challenges above emulation does have several benefits, which is why they remain part of the digital preservation toolkit
- They are a good choice when the look and feel of digital content is important to retain, such as with digital artworks
- They offer a user experience closer to the original use context than seeing a list of files or browsing content in a contemporary computing environment.
- Emulators are also useful in content appraisal. In order to determine if a collection or file is to be assessed for enduring value, curators must understand what the file content is and what it means overall.

Slide 5: Involvement with FCoP:

- An opportunity to engage with emulation and software preservation on a community level presented itself in Jan. 2018 through the call for

proposals for the **Fostering Communities of Practice: Software Preservation and Emulation in Libraries, Archives and Museums,** or the FCoP. Institute for Library and Museum Services [IMLS grant RE-95-17-0058-17] subproject

- The focus of our involvement in FCoP is in preserving, improving discovery of and providing access to files created by contemporary music composers. These collections are stewarded by the Sousa Archives and Center for American Music.
- We are particularly interested in further investigation and development of an emulated/virtual environments where these titles can run in as close to a native environment as possible. The collection curator, Scott Schwartz, interest in emulation is in presenting the files in as close as we can get to the creators' working context.
- In most of the audio production or composition context, recorded output is not enough to demonstrate the creative context. Scott equates having born-digital production files to having access to a composer's notebook where a researcher may gain additional information about what creative choices were made when composing or producing audio works.
- From a service point of view we are interested in scaling this work to meet the needs of future collections of composers' and other born-digital collections with consideration of available resources.
- The FCoP project has facilitated prioritizing emulation and software preservation, particularly within the broad landscape of institutional digital preservation activities where there are often competing priorities for time and attention.

Slide 6: Collections identified for this project:

- Initially centered around born-digital collections of three Illinois composers. Each collection presents curation challenges and different types of information provided about the respective collection items.

- The creation dates within the collections span from 1992 – 2012, representing a significant expanse of time in terms of technological development and software versions.
    - Michael Manion:
        - The born-digital content from the Michael Manion Music and Papers were recovered from his Macintosh PowerBook 3400c, manufactured early 1997.
        - Its operating system is Mac OS 8.6. Software of note within this collection are composition and arrangement related Band-in-a-Box and music notation program Finale. Both software titles are versions circa late 90s.
        - Little information about how to approach curation or which files Michael created. A significant amount of curation work is required to identify Manion's files and to provide access to them.
    - Peter Micahlove:
        - Born-digital content recovered from a laptop running Windows 7.
        - This collection arrived with a file inventory created by Peter Michalove which provided guidance for focused curation efforts as we have a roadmap of files of interest rather than sifting through the entire computer file system. This inventory is a useful document to use in appraising the collection.
        - We initially used this laptop as a use case for accessing disk imaged via a virtual environment. The outcome was not especially successful as setting up the virtual environment demonstrated the resource intensiveness required for such an endeavor. The computer and subsequent disk image contained malware which caused antivirus

alerts when the disk image was mounted and still required significant curation before we could allow user access to the VM.

- Scott Wyatt:
  - Receive a file transfer of Pro Tools session and audio output files from Wyatt
  - Compared to the other two collections the creator is still alive and available to ask question of as we curate this collection.

- The biggest challenge with this collection is running the Pro Tools digital audio production workstation environment and researching the proprietary file properties and associated dependencies.
- I quickly realized that this scope of work was too ambitious for the time available within the project cycle or found that some instances would not be well served through emulation.
- For example, running ProTools requires that we use a hardware authentication method. Also, the ProTools files found within the Wyatt collection can be rendered in the current version of ProTools. We've thus decided to, at present, provide unemulated reading room access to the Wyatt collection and emulate at a later time should it make sense to do so. We've documented important information about the ProTools files to better facilitate emulation or other modes of access.
- Our focus has been on the Manion computer; we have been working within EaaSI on the Manion computing environment. Currently, we have an EaaSI environment which was created from the forensic .E01 disk image format. We are currently using this disk image and the EaaSI environment as a processing area – reviewing the files in their native software, reviewing the software installed within the creation environment and identifying where there are hardware dependencies and other settings idiosyncratic to this composer, but of classes of activity which might be generalizable enough to considered significant when assessing other composers files.

Slide Seven: Scaling and Documenting Curation Efforts:

Much of my work within the FCoP project has been centered on formalizing and scaling workflows. Most of the methods used within the scope of the project were in various stages of development prior to the project. However, involvement in the project has emphasized to me the importance of these steps to continue to build access capabilities and scaling the work beyond myself.

Documentation:

1. **Documenting Locally Significant Formats:**
    a. **Digital Content Format Registry:**
        i. The resources required to curate legacy content in order to render the files is considerable. Curating this content to full functionality requires software, knowledge on how to run the software and associated operating system environment, it requires patience to discover and resolve errors which arise from software and file dependencies (such as specific fonts linked to a document) or making the decisions about what errors are acceptable (or desirable to maintain) and documenting choices made.
        ii. Available staff with the time and knowledge to undertake this work and support it programmatically is limited
        iii. We've have undertaken several efforts to build capacity for improving access through file rendering. These efforts are often designed to build upon current digital preservation practice, demonstrating one aspect of the iterative nature of digital preservation
        iv. One strategy is in building the Digital Content Format Registry. This is a tool developed in by our preservation librarian and the Medusa software development team. The research focus of this tool is to documenting local knowledge gathered about how to identify and render challenging file formats – particularly formats that

present challenges including being associated with a specific version of proprietary software.

v. These formats are often also complex, with dependencies upon hardware, software and other files to render successfully.

vi. Information about these formats also tends to be weak or non-existent in international or large-scale file format identification tools such as FITS or DROID (tools commonly used to automate file format validation and identification) due to the challenges associated with these formats.

vii. Strategies used in determining how to access challenging file formats include locating software to run the files and, in some cases, using available emulators to run single files.

2. **Cataloging Software:**
    a. About Hembrough collection: large donation of software circa late-80s/mid-90s. Mostly Windows-based business productivity software titles. Some in original packaging with documentation.
    b. Others are copies of software (utilities; mostly received via mail order. Small companies that may have created one-off applcations
    c. G:\Preservation&Conservation\Departments\Digital Preservation\Digital Preservation and Processing Tools\HembroughSoftwareDonation
    d. About Hembrough collection: large donation of software circa late-80s/mid-90s. Mostly Windows-based business productivity software titles. Some in original packaging with documentation.
    e. Others are copies of software (utilities; mostly received via mail order. Small companies that may have created one-off applcations.)
    f. Various students have been working on cataloging these titles. Based data fields partially off of Cabrinety-NIST collection organization at Stanford (https://library.stanford.edu/projects/cabrinety-nist-project).
    g. Have only cataloged. Next steps include review for import into EaaSI node; reviewing catalog data to better align with shared findability standards. e.g. Wikidata

Slide Eight: Scaling and sharing the work: Curation Workflows:

- Emulation and access efforts have underscored the importance of gathering information and permissions about collections as early as possible. To this end, I've been reviewing and will be focusing continued effort on outreach to collections managers on what they can do early on in the acquisition process
    - Addendum to deed of gift for electronic records
    - Guidelines for Donating Digital Materials: This document provides introductory information on points to consider when donating computer based files
    - Acquisitions documents/guidance: preservation appraisal guidelines
- **Creating workflows for testing software:**
    - Another part of scaling the work is creating working methods for relaying how to complete complex work. For example, quickly imparting enough information to a student worker or other not typically doing work within the area of digital preservation to install legacy software is a challenge. But the responsibility for this work must be shared in order to make efforts scalable.
    - To this end I'm in the process of developing and testing docs and workflows for:
        - Installing and testing legacy software – the workflow breaks out steps for information to collect and is also intended to provide steps for information to capture when there is a roadblock in the installation. I actively discourage emulation as a foregone conclusion – instead, having staff work through the steps, document what they observe and also how long it take them to complete each activity.
        - Also working on expanding disk imaging workflows to QA test software disk images
        - Presently, moving toward creating a software preservation collection within our digital repository. Considering collection development policies to inform what is collected, how the collections are structured and how to encompass manuals and contextual information about the software.
        - Workflow for emulation: I'm purposefully building out access workflows to encompass emulation as a method for access and not an end-goal. Getting to the point of emulation has largely used workflows that I'd previously had in place but are slightly expanded or reviewed to allow for emulation as an access solution.
        - I'll discuss conclusions and future work later in this talk. As emphasized, this is an ongoing project.
        - Now onto Lauren!