Preservation Week Talk: Tuesday, April 23rd, 1-2pm, RM 106

1. **Title slide**
2. **Intro:**
    a. Thank you all for coming to this Preservation Week event.
        i. About preservation week: from ALCTS http://www.ala.org/alcts/preservationweek/about
            1. An initiative of the ALA vision, the Association for Library Collections and technical services designed to encourage awareness of preservation issues
        ii. Special thank you to Miriam Centeno for coordinating the event
        iii. There are other events each day this week. Here's the link to the schedule and more info: https://wordpress.library.illinois.edu/staff/preservation/services/education_training/preservation-week-2019/

3. **Outline**
    1. Broad overview of select digital preservation activities
    2. About emulation
    3. About the Fostering a Community of Practice: Software Preservation and Emulation in Libraries, Archives and Museum project
    4. Our role in and ongoing activities related to FCoP

4. **Broad overview of digital preservation at Illinois:**
    a. One service point is the born-digital reformatting lab where collections content is migrated from fragile computer media such as floppy disks and older internal hard drives. The content is migrated using digital forensics techniques to maintain provenance. After recovery and processing, the content selected for long-term preservation is then moved into our digital preservation repository.
    b. The content on these obsolete carriers often is of the same era, or older, as the carriers
    c. The collection content often represents a 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.

d. 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. With the contents stored in Medusa have access to the "bits on the disk" to continue refining and developing our programmatic digital preservation efforts.
e. Most of the collections I work with are from special collections units like University Archives, Sousa Archives, Illinois History and Lincoln Collections, and the RBML.

5. Software Dependency

a. These born-digital collections objects are often dependent upon obsolete software, or legacy versions of contemporary software titles which likely has changed greatly in supported functionality between releases.
b. 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.
c. 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.

6. File Format Registry

a. 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 and documenting those errors.

b. Available staff with the time and knowledge to undertake this work and support it programmatically is limited

c. 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

   i. One strategy is in building the File Format Registry. This is a tool developed in part by Kyle Rimkus 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.

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

   iii. 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.

   iv. 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.

7. What is Emulation?

a. I mentioned that we have used emulators to access single files as part of the file format registry data collection, but you may not know what an emulator is.

b. " Emulation – combines software and hardware to reproduce in all essential characteristics the performance of another computer of a different design, allowing programs or media designed for a particular environment to operate in a different, usually newer environment"

https://dpworkshop.org/dpm-eng/terminology/strategies.html (Digital Preservation Management - Digital Preservation Strategies)

c. That is, an emulator is software that mimics the behavior of another computing environment and is used to access software and digital files which require access to obsolete technological environments in order to run.

d. Emulation strategies focus on hardware and software environment recreation rather than transforming the digital object. The original file remains unmodified; it is the computing environment that changes.

e. Development of emulators and use has been strong in some hobbyest user communities such as gaming

f. One example: image 3 is a Screenshot of DOSBox v0.74 interface **DOSBox** is an emulator program which emulates an IBM PC compatible computer running a DOS operating system. It's free software that has elements of both software and hardware emulation.

8. Emulation in Practice:

a. 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

b. Widespread and scalable use limited as there are steep resource barriers to entry

c. Emulation often require significant technological knowledge and administrative resources to research, develop and implement solutions

d. 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.

e. Efforts centered around specialized implementations are often too resource intensive and not scalable enough for practical implementation

   i. For example, the Salman Rushdie emulation project at Emory where four of the author Salman Rushdie's Macintosh

computers were emulated is often looked to as an example of emulation used in digital preservation practice.

ii. However, this example does not demonstrate scalability and indeed required many grant-funded resources to make it happen.

iii. Those presently responsible for stewarding these collections have recognized that the grant-funded approach employed at the time was not sustainable and are seeking ways to make future efforts scalable and manageable.

iv. Notably, upon review and reassessment of this project they also learned that the need to document everything the software engineers do as their work is just as ephemeral as the born digital information they wished to preserve.

9. Benefits of emulation:

a. Despite the challenges above emulation does have several benefits, which is why they remain part of the digital preservation toolkit

b. They are a good choice when the look and feel of digital content is important to retain, such as with digital artworks

c. 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.

d. 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.

e. As we learned during the file format registry project sometimes an environment to run legacy software is needed to determine what a file contains so those appraisal decisions can be made. In some cases a hex editor can be used to review text stored in an obsolete file format, but the experience of the file is lost.

f. Reviewing non-text-based files in a hex editor is often insufficient as a “performance” of file better represents the significant aspects of the file.

g. Emulation (indeed most digital preservation strategies) is often used in combination with other digital preservation strategies

i. For example, we might be able to migrate a file format to a contemporary version of the format; however, getting to that

point may require an emulator to run earlier version of the software to facilitate exporting to a version. The emulator can thus act as a bridge technology even though the primary strategy used is migration.

ii. Emulators can also help us evaluate information loss. As file formats are migrated, they're especially prone to information loss. If we can evaluate that file in an environment closer to its native creation environment and compare how it performs, we can assess if the significant properties of the file has been retained or assess and record information loss.

10. Emulation in relation to software preservation:

a. Sometimes implicit when considering emulation is the need for keeping software executables or source code and associated documentation available for access.
b. Our digital cultural record depends upon software preservation to retain and render software-dependent digital objects. It is important as we develop and improve strategies for content access. Each one of our digital files depends upon some level of software mediation in order to be accessible.
c. Successful emulation environments are heavily dependent upon software availability.
d. In order to run software in a software-based emulator you need access to the executables or source code of the software you want to run. You also need adjacent technical software such as the operating system and hardware drivers.

11. **An opportunity to develop local practice collaboratively:**

a. In digital preservation services we have had a few forays into applying emulation. However, we too encountered common roadblocks of lack of resources and scalable solutions.
b. 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

c. What it is: The FCoP is an Institute for Library and Museum Services [IMLS grant RE-95-17-0058-17] subproject. A cohort of six organizations were selected to undertake software preservation and emulation projects to establish a community of practice in software preservation and emulation within libraries, archives and museums.
d. What problem is it trying to solve?
    i. Sharing in the development and implementation software preservation and emulation solutions and practices in a scalable and sustainable manner through developing a cohort of users. The cohort model is intended to build on shared capacity and community and lower the barrier to entry to software preservation and emulation solutions.
e. **Who is involved?**
    i. PI: Zach Vowell, Cal Poly State University
    Project Coordinator: Jessica Meyerson, Educopia Institute
    ii. Cohort Institutions:
        1. University of Illinois Urbana-Champaign
        2. University of Viriginia
        3. University of Arizona
        4. Georgia Tech
        5. Guggeneheim Museum
        6. Living Computers: Museums + Labs
    iii. Local project team:
        1. Tracy Popp (project lead)
        2. Kyle Rimkus
        3. Seth Robbins
        4. Karl Germeck

12.About our Project:

a. We are interested 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.
b. 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. Scott

Schwartz' interest in emulation is in presenting the files in as close as we can get to the creators' working context.

c. 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.
d. 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.
e. Collections identified for this project:
    i. 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.
    ii. The creation dates within the collections span from 1992 – 2012, representing a significant expanse of time in terms of technological development and software versions.
        1. Michael Manion:
            a. The born-digital content from the Michael Manion Music and Papers were recovered from his Macintosh PowerBook 3400c, manufactured early 1997.
            b. 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.
            c. 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.
        2. Peter Micahlove:
            a. Born-digital content recovered from a laptop running Windows 7.

b. 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.
c. 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.

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

13.Project timeline:

a. Proposal accepted late winter 2018.
b. Early organization activities and project ramp-up in late spring to end of July 2018
c. Projects start in earnest at in-person kick-off meeting held at the Computer History Museum in Mountain View, CA August 1-3, 2018

d. Core project activity from August 2018 through November 2019. So, we're presently in the middle of the project and heading into the last six months.
e. Core research and investigation areas:
    i. Legal
        1. Association of Research Libraries lead Code of Best Practices in Fair Use for Software Preservation: intended to provide institutions clear guidance on the legality of archiving software in order to ensure continued access to digital files of all kinds
    ii. Metadata
    iii. Technical preservation
    iv. Access challenges
    v. Knowledge development
    vi. Outreach and information sharing:
        1. Outreach and the overall grant timeline haven't quite aligned. Our time is compressed in order to attend key conferences within the IMLS grant timeline.
        2. We're already preparing for our first round of workshops and information sharing: I'll be at the Society of American Archivists' 2019 conference participating in a day long workshop.
        3. A subset of the cohort (of which I'm included) has sent a panel proposal to iPRES 2019.
    vii. Experimenting and testing emulation software
        1. Emulation testing sandbox provided through concurrent research project in the Software Preservation Network's portfolio:
            a. Scaling Emulation and Software Preservation Infrastructure (EaaSI)

14.Scaling Emulation and Software Preservation Infrastructure (EaaSI)

- EaaSI is a concurrent grant project under the Software Preservation Network administrative umbrella
- The EaaSI program builds on previous work to apply the Emulation-as-a-Service(EaaS) framework for access and use of preserved software and digital objects and is focused on scaling the

technological framework necessary for multiple institutions to configure, share, and access software and configured environments. It's lead by the Digital Preservation services team at Yale University Library.

- The EaaSI network includes access to configured software environments, that is, a representation of a technology stack that includes the operating system, configuration of specific OS settings, installation of drivers appropriates to the software applications of the same computing era. It is accessed via a web interface. AS of this writing it is in beta release
- The EaaSI user handbook is available here: https://eaasi.gitlab.io/eaasi_user_handbook/guide/introduction.html
- The FCoP cohort is "kicking the tires" and assisting in the development of EaaSI as we work through our projects. We have been working closely with the technical developers, submitting error reports and feature requests.
- Institutions make decisions about which software and what versions they need to use based on their collection needs.
- The EaaSI team reports back about what features they've implemented; problems addressed and help us install operating system environments on request.

15.Local FCoP activities:

a. Lots of activities happening concurrently within the cohort and other related projects which affects our local timeline
b. What stage we're in:
    i. I'm working on outreach at the cohort level which is influencing and compressing the overall project timeline as we're starting outreach as early late July 2019 at SAA 2019.
    ii. As I said we're still in the middle core work of the project.
    iii. Locally this means the team is deep into curation work.
    iv. As we've spent more time with the local work curation work as well as the intensive cohort and research level work our efforts have been scaled back to addressing concerns in the the Manion collection. I made this decision as it needs the most curation work and represents a computing environment that is most amenable to running in the EaaSI environment.

v. Presently, we do have an emulated version of his laptop running in EaaSI.
vi. The emulated environment was generated from forensic disk image and we can access it in EaasI (show here):
    1. https://illinois.softwarepreservationnetwork.org
    2. Walk through environment, run Manion to demonstrate the interface

16. **About this environment**
    i. This demonstration of the emulated environment generated from a forensic disk image represents an important milestone, it only represents a small portion of the project work. To me, this stage highlights more questions than providing definitive answers. In addition to the technical work I am asking questions and drafting guidance documents related to workflow, resources and scaling efforts (not as fun nor as sexy as demo'ing the emulator)
    ii. I consider this environment useful in appraisal but not for researcher access. It is helpful as we perform technical file and software appraisal and document information about the software environment and make decisions on how to provide access to this content.
    iii. But the content has not been processed – it is a representation of the computer hard disk drive as recovered from the laptop. Files have not been scanned for sensitive information.
    iv. There are still technical issues. At present, the audio playback is not working. This will require further troubleshooting to determine what is preventing audio playback.
    v. Moving into a scalable and service-level implementation require significant curation work which must be done with engagement from content curators, making decisions based on preservation priorities, documenting what we've done to continue to build digital preservation capabilities, and making the work visible in order to share with others interested in undertaking software preservation and emulation efforts and to illustrate the workflows.

17.Archivists and Curators Engaging:

a. Work to date has underlined the need for archivists and curators to:
    i. Gather as much information as they can from the outset about the creation context and files of interest if we are to provide access to collections via emulation.
        1. They're the first point of contact with the collection donor
        2. What information they gather can influence preservation outcomes
    ii. Document use and have the creator or donor walk through the software interface if possible, particularly if the software environment is complex and has custom settings which are key to rendering the files
        1. Ask about software use – versions used, what was created, what other tools were used with the software.
            a. A great example is a video created by SACAM talking to Scott Wyatt about his use of Pro Tools and having Scott walk through use of the program
        2. Identify files of value at acquisition
            a. Can the donor document the file structure or provide a file inventory?
    iii. In encouraging donors to make file inventories or identifying "files of value" we Don't necessarily want to encourage transfer to a secondary storage medium. That can lead to important metadata such as creation dates may be lost. If in doubt engage digital preservation specialists early in the acquisition process to discuss strategies.
    iv. Gain clear permissions on use and access
        1. Address issues related to digital preservation actions and access in the acquisition stage and in deed of gift.
        2. Possible transfer of software as part of the donation

18.Selection and Prioritization:

a. Although the EaaSI project addresses and manages some of the challenges of providing emulation services, it is still resource intensive. Implementing this as a service still require establishing criteria for emulation use as an access strategy and.

b. The team assembled for this project is only temporary – I'm the only permanent staff on the project team and I have many other job duties to address. As with most preservation projects prioritization and balancing resources is required.

c. Part of this management includes developing project plans for candidate collections

d. Some criteria and questions to consider are:
    i. What makes a good candidate for emulation services for user access and for appraisal?
    ii. and at what level will collections be emulated
        1. meaning will just a class of files be emulated in a general operating system environment (this follow
        2. or will a specific operating environment be created to represent one collection
    iii. How to prioritize which collection receive this type of access.
    iv. Gauging if there is enough information about the software, dependencies and files required to render the files or to create a reasonable facsimile of the creator's computing environment

19. Making the work Visible and Guiding others:

a. Making this scalable also means
    i. engaging and training others within the organization to share the workload
    ii. I'm articulating and documenting the assessment work and processes that I undertake in order to make the work visible and for others to use it in training documentation.
    iii. In drafting these documents and having other team members walk through them, illuminates work load and common roadblocks and tests sharing the workload, and the training methods.
    iv. Identifying responsibility such as who will do the work of:
        1. Procuring software
        2. Installing software
        3. Knowing how to use the software and to what extent
        4. Support and maintenance of:
            a. Software install files and associated documentation

v. Outreach is important for information sharing within the library and to garner support
vi. As part of the FCoP project I am creating documents to share with other external to U of I who may be interested in undertaking software preservation and emulation efforts.

20. Future activities:

a. User testing of the EaaSI environment – we will be working with Scott Schwartz to identify a group of researchers to test and comment on the EaaSI portal for collections research access.
b. Using the EaaSI tool to create migration pathways
    i. Investigating creating pathways that do not completely rely on emulation for access
    ii. For example, thinking about other access methods or versions of files we may migrate to mitigate encountering future emulation access
    iii. Being a software solution emulation is also subject to obsolescence. While we can still render a file and are actively engaged in understanding a file format and its construction seems to be a good time to consider how else we might preserve key information from and about the file and contents in a way that isn't reliant on emulation.
    iv. As noted, I do not yet know what we will have access to in terms of emulation after this project or how we will access EaaSI.
c. Developing a more formalized software preservation practice
    i. creating local storage in our digital preservation repository, Medusa, for our collection of software in use. This may require developing a collection development policy and associated collection management guidelines.
    ii. Further development of software metadata and inventory sharing to share with others
d. Determining if we will have continued access to EaaSi?
    i. It is unclear how access to EaaSI will be facilitated at the end of this project. The University Library is currently a member of the nascent Software Preservation Network. Through this work

I am hoping to learn more about how they plan to support access to EaaSI (or not)

e. Continue working with the Software Preservation Network to help guide software preservation in practice at a consortia level. As part of our SPN membership I am a voice in helping shape the professional organization and services models.

21. Resources

22. Image Sources:

23. Thank You and Questions