2019-11-20, 4p 308 Main Library – Data Interest Group

Guiding questions for presentation portion:

- What is the question to be answered?
- With what data?
- 15 minutes MAX
- At least one slide

2. **Digital Preservation @ University Library**
    a. Today, I'm going to talk about concepts related to digital curation and stewardship and maintenance activities intended to keep digital data accessible.
    b. I'm going focus on my involvement with the Fostering Communities of Practice: Emulation and Software Preservation in Libraries, Archives and Museums
    c. About me: Digital Preservation Coordinator
        i. Work centers around preserving and making accessible digital (particularly born-digital) collections material
    d. 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.
    e. 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.
    f. 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
    g. With the contents stored **in Medusa have access to the "bits on the disk"** to continue refining and developing our programmatic digital preservation efforts.

3. **Fostering a Community of Practice project**
    a. **Question to be answered**: Is emulation a viable and scalable digital preservation tool for use in libraries, archives, and museums?

b. **An opportunity to develop local practice collaboratively:**
    i. In digital preservation services we have had a few forays into applying emulation. However, we encountered common roadblocks of lack of resources and scalable solutions.
c. 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 subproject
d. What it is: The FCoP is an Institute for Library and Museum Services 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.
e. 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.
    ii. The cohort model is intended to build on shared capacity and community and lower the barrier to entry to software preservation and emulation solutions.
    iii. 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

4. **Scaling Emulation and Software Preservation Infrastructure (EaaSI)**
    a. EaaSI is a concurrent grant project under the Software Preservation Network administrative umbrella
    b. 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.

c. 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
d. The EaaSI user handbook is available here: https://eaasi.gitlab.io/eaasi_user_handbook/guide/introduction.html
e. 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.
f. Institutions make decisions about which software and what versions they need to use based on their collection needs.
g. The EaaSI team reports back about what features they've implemented; problems addressed and help us install operating system environments on request.

5. **What is Emulation?**
    a. What is it?
    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. 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
    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.
g. Widespread and scalable use limited as there are steep resource barriers to entry
h. Emulation often require significant technological knowledge and administrative resources to research, develop and implement solutions
i. Sometimes implicit when considering emulation is the need for keeping software executables or source code and associated documentation available for access.
j. 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.

6. **Emulation Benefits**

a. **Why use it?**
    i. Useful when "look and feel" or close approximation is important (digital art, etc.)
    ii. Seeing the content "in situ" – modifying the computing environment, not the content – potentially more engaging than just looking at a list of files in a spreadsheet
    iii. Can be used in archival processing
        1. Can load a legacy disk image and see contents in context to gain a better understanding a creator's computing environment, what applications they were using, how they were storing their files, reviewing files that may not have file extensions when exported from the disk image (a common Mac problem)
        2. Extract content from the emulation environment in a what I'm calling a subtractive processing method
        3. Or create an environment (OS, application, files stack) and use this as a processed collection environment (additive method)
    iv. Can use it as a QC tool
        1. May encounter issues when data is migrated – for example, numerical data may not retain precision (or be represented correctly) when migrated from one spreadsheet application to another
        2. Can create a computing environment where one can run a version used to create the software (or close to) and verify migrated output
b. Cons: 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.
c. Emulation often require significant technological knowledge and administrative resources to research, develop and implement solutions

7. **Illinois' FCoP 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. 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.
c. 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
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.
    i. 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.
    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.

i. Scaling the work includes developing workflows to accommodate future collections

8. **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've starting outreach as early late August 2019 at SAA 2019.
        ii. We're still in the core work of the project but heading into the final phase
        iii. Locally this means I'm is deep into curation work, documenting and testing workflows
        iv. Efforts have been scaled back to primarily addressing concerns in the the Manion collection although I have done some work with Michalove.
        v. 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.
        vi. Presently, we do have an emulated version of his laptop running in EaaSI.
9. **About the Emulated Environment**
        i. Manion's emulated environment was generated from forensic disk image and we can access it in EaasI
        ii. 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)
        iii. 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.

iv. 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.
v. 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.
vi. 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.

10. **EaaSI in Action**
    a. Screen capture of launching a Windows 7 environment in EaaSI. The primary goal of this environment is to launch a digital humanities project that was stored on optical disk. The project is essentially a website dependent upon the Flash multimedia software browser plugin to render. Flash was once the dominant platform for online multimedia content, but has slowly been abandoned in favor of less resource intensive and more secure applications. Notably, Apple did not build in support for Flash in their phones, which signaled a death knell for the platform. The Flash Player has an official end-of-life at the end of 2020
    b. Required that I find a browser that would allow me to install a Flash plugin. Fortunately, Mozilla has retained an archive of most of their previous versions of Firefox. A Google search regarding which last version of Firefox that supported Flash content pointed to the version that I needed which would run in Windows 7. I was also able to download a standalone version of the Adobe Flash Player to install in the Windows 7 environment. With this software in place I was able

to launch the project from a virtual CD and render the website and associated video content.

11. **Scaling and Documenting Curation**

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: 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.

12. **Engaging Archivists and Curators**
    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.

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

b. Addendum to deed of gift for electronic records
c. Guidelines for Donating Digital Materials: This document provides introductory information on points to consider when donating computer based files
d. Acquisitions documents/guidance: preservation appraisal guidelines

13. **Making the Work Visible and Sharing Experiences**
    a. Emulation isn't a magic box – levels of rendering success will depend upon the complexity of the creator's computing environments and the work undertaken to analyze the content
    b. 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.

14. **Future Work**
    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.

15. **Select Resources**
16. **Thank you!**