

Carter Garde Hulinsky. Title. A Master's Paper for the M.S. in L.S degree. May, 2023.
63 pages Advisor: Helen Tibbo

This paper discusses the challenges of implementing the Open Archival Information System (OAIS) reference model for born-digital materials in digital preservation.

Although the OAIS model has been globally recognized for its universal terminologies and conceptual standards, it offers little guidance in terms of tangible implementation.

Consequently, archivists have created various methods and tools for OAIS-compliant digital archival preservation workflows. This paper presents a project at Duke University Medical Center Archives, which aims to enhance the repository's current Electronic Records Processing Guide using the digital materials from two recent accessions. The revised guide will be tested and developed, utilizing open-source digital forensic tools to process electronic records for ingest into the repository's OAIS-compliant integrated archives management system. The outcomes of this project will provide increased stability and efficiency in processing a larger volume of digital materials.

Headings:

Digital Preservation

Electronic records

Archives

FROM OVERWHELMED TO OVERCOMING: DEVELOPING A PRE-INGEST
PROCESSING MANUAL FOR BORN DIGITAL CONTENT

by
Carter Garde Hulinsky

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Library Science.

Chapel Hill, North Carolina

May 2023

Approved by

Helen Tibbo

Table of Contents

INTRODUCTION	3
LITERATURE REVIEW	5
1.1 Introduction	5
1.2 Background	6
1.3 Components of Pre-Ingest Processing.....	8
1.3.1 Accessioning and tracking	8
1.3.2 File integrity checking and fixity	9
1.3.3 Virus scanning and malware detection	11
1.3.4 File format assessment.....	12
1.3.5 Appraisal and selection	13
1.3.6 Sensitivity Review	16
1.3.7 Metadata creation and management	17
1.4 Professional standards and guidelines for Pre-Ingest Processing	18
1.4.1 Open Archival Information System	18
1.4.2 Digital Processing Workstation.....	18
1.5 Challenges of effective pre-ingest processing	20
1.6 Conclusion.....	21
METHODOLOGY:	23
1.1 Positionality Statement:.....	23
1.2 Project Background	24
1.3 Organization Description: Duke University Medical Center Archives.....	26
1.4 AXAEM: An Integrated Records Management System	27
1.5 An Important Consideration: HIPPA & FERPA.....	29
1.6 Current Electronic Records Processing Guide Description	30
PROJECT OUTCOME:.....	33
1.1 Pre-Ingest Processing Manual for Digital Content	33
1.2 Introduction	34
1.3 Linux Basics and Helpful Hints	35
1.4 System Update.....	36
1.5 Wired Connection (Disconnect/Connect).....	36
1.6 File Fixity	37
1.6.1 Generating a Checksum Report:.....	37

1.6.2	Executing a Checksum Check:	38
1.7	Virus Scan using Clam AV	39
1.8	Creating Copies, Master Files, Ingest Files, and Working Files	40
1.8.1	Multiple copies, their purpose, and where to store them?	40
1.9	Reviewing Files.....	40
	What should we keep?	40
1.9.1	File Type Identification and Count.....	41
1.9.2	List All Files Ordered by Size	42
1.9.3	List All Files by Last Modified Date	43
1.9.4	Identifying and Removing Zero-Byte Files?	43
1.9.5	Counting Files	44
1.9.6	Calculating directory size (bytes).....	44
1.10	What about duplicates? Using fdupes	45
1.11	What about PII? Sensitivity Review using Bulk Extractor and Risk Assessment	46
1.12	Renaming Files	49
	<i>ASSESSMENT</i>	<i>50</i>
	<i>DISCUSSION:</i>.....	<i>52</i>
1.1	Key Stakeholders.....	52
1.2	Steps Taken	53
1.3	Challenges Encountered	54
	<i>CONCLUSION:</i>	<i>57</i>
	<i>WORKS CITED:</i>.....	<i>59</i>

INTRODUCTION

For decades archivists have understood the necessity of preserving born-digital materials but grappled with how to safeguard authenticity, manage descriptive metadata, and provide access to those with information needs. Conceptual workflows, such as the Open Archival Information System (OAIS) reference model, were created by working groups to address the unique challenges posed by digital preservation. The OAIS model is internationally recognized for defining universal terminologies and providing conceptual standards at a macroscopic level. The global recognition of OAIS has stimulated discussion, implemented processes, and served as a basis for the evaluation of digital preservation workflows. While the OAIS model has been pervasive in the field of digital preservation, the model by its very design offers little guidance in the tangible sense to its implementation. As a result, archivists have actuated many methods and tools for OAIS-compliant digital archival preservation workflows. The documentation of these methods, particularly with born-digital archival processing, remains inadequate in current scholarly literature. By processing, I mean the preparation of archival materials for access by the appraisal, arrangement, description, and review of sensitive information.

This paper will describe a project at Duke University Medical Center Archives involving the enhancement of the repository's current *Electronic Records Processing Guide* with a supplemental pre ingest processing manual for born digital content. The additional manual will be developed and tested using the digital materials from two

recent digital accessions to the Garnett H. Kelsoe Laboratory Notebooks and the Physician Assistant Program Records. The outcomes of the revised guide will provide for increase stability in the transfer of electronic records from digital carriers and efficiency in processing a larger volume of digital materials by utilizing open-source digital forensic tools to process them for ingest into the repository's OAIS-compliant integrated archives management system.

LITERATURE REVIEW

1.1 Introduction

Archivists and digital preservation experts are facing new challenges due to the growing prevalence of born-digital content in recent years. Born-digital content refers to materials that are created and managed in digital form, including digital photographs, documents, harvested web content, manuscripts, electronic records, static and dynamic data sets, art, and digital media publications (Erway, 2010). Unlike physical materials, born digital content requires specialized tools and techniques for preservation and access. Goldman (2011) cautions that waiting for a perfect, affordable, and all-encompassing solution for born-digital content preservation is impractical. The digital content in our possession is already deteriorating, and the digital universe is rapidly expanding. One of the critical stages in managing born digital content is pre-ingest processing, which involves preparing digital files for ingest as a submission information package (SIP) into a digital asset management system which facilitates their preservation and access. Pre-ingest processing refers to a set of activities by digital archivist which involve the migration, appraisal, organization, and preparation of digital content to facilitate a successful packaging of materials for ingest. These activities include accessioning and tracking, file integrity checking and fixity, virus scanning and malware detection, file format assessment, appraisal and selection, sensitivity review, metadata creation and management. Pre-ingest processing is critical to ensuring the long-term preservation and

access to digital content. Without proper pre-ingest processing, digital archives and repositories may face a variety of challenges, including format obsolescence, data loss, and security breaches (Erway, 2010; Baucom, 2019) Additionally, accurate and comprehensive metadata is critical to the discovery and use of digital content. Pre-ingest processing also allows for effective appraisal and selection of digital content, which is essential for ensuring that digital archives and repositories are sustainable and meet the information needs of their users.

Given the importance of pre-ingest processing for managing born digital content, there is a need for a comprehensive understanding of its components, best practices, and challenges. The purpose of this literature review is to examine the components of a pre-ingest processing manual for archival digital files. The review will provide an overview of the importance of pre-ingest processing in digital archives, the challenges faced by archivists in managing digital content, and the best practices for pre-ingest processing. The review will also highlight professional standards and guidelines for effective pre-ingest processing and will present case studies and examples of successful pre-ingest processing. Ultimately, the review aims to provide recommendations for future research and development in pre-ingest processing for born digital content, to enhance the preservation and access of digital materials.

1.2 Background

The concept of digital archives can be traced back to the early days of computing in the 1960s and 1970s when organizations began to use computers to store and manage electronic data. However, it was not until the 1980s and 1990s that digital archives began to emerge as a distinct field of study and practice. During this period, advances in

personal computer technology and the growth of the internet led to a proliferation of digital content, including websites, email messages, and multimedia files (Baucom, 2019). As the volume of digital content grew, so did the need for effective strategies for managing and preserving it.

One of the key challenges facing digital archives was the issue of file format obsolescence. Unlike physical materials, digital files are dependent on software and hardware to access and use them. As software and hardware evolve, older file formats may become obsolete, rendering digital content inaccessible. Another challenge is the occurrence of bit rot, which refers to the decay or corruption of the bit streams that make up digital content (Baucom, 2019). Archivists have responded to the challenge of preserving digital files by creating plans to convert them to newer, more durable formats. The pre-ingest processing stage has emerged as a vital step in the digital archives workflow, where digital files are evaluated and converted to ensure their continued preservation and accessibility. Having established workflows provides clear guidance and examples for organizations to use their current processes to identify areas for improvement, set goals for growth and advancement, and request additional resources while restructuring roles and relationships as necessary to enhance their ability to preserve born-digital materials (Chassanoff and Post, 2020).

Since the early 2000s, a variety of tools and techniques have been developed to support pre-ingest processing, including file format identification and characterization tools, metadata creation and management tools, and checksum verification tools (Walsh, 2017). Additionally, several digital preservation communities and organizations, such as the Digital Preservation Coalition and the Open Preservation Foundation, emerged to

promote best practices and standards for digital preservation. Today, pre-ingest processing remains a critical stage in the digital archives workflow, with archivists and digital preservation professionals using a variety of tools and techniques to ensure the long-term preservation and access of digital content. While the field of digital archives and preservation continues to evolve, pre-ingest processing remains an essential component of effective digital preservation strategies.

1.3 Components of Pre-Ingest Processing

1.3.1 Accessioning and tracking

Accessioning refers to the process of accepting digital content into the archive and assigning it a unique identifier to track its movement throughout the preservation process. This identifier helps to ensure that the digital content is easily traceable and identifiable within the archive. According to Woods & Lee (2012), captured disk images do not preserve any additional metadata regarding the process or supporting actions that were performed during acquisition. Therefore, the digital archivist needs to record information about the digital content during the accessioning process. This should include information about the creator of the content, the date of creation, and any other associated metadata. This step is necessary to provide context and ensure that the digital content remains accessible and usable over time.

Furthermore, a survey conducted by DeRidder and Helms in 2016 found that the top collected metadata for digital files were file dates, file types, size, and checksums. These metadata elements can provide valuable information about the digital content and help in the management of digital collections. File dates can help to establish the creation and modification dates of a file, while file types can provide information about the format

of the digital content. The file size can indicate the amount of storage required for preservation, while checksums can be used for file integrity checking to ensure that the file has not been altered.

Chassanoff & Post (2020) recommend that in addition to addressing permissions and identifying any sensitive information, archivists should collaborate closely with the donor to gather information about how the digital content was created and maintained. This process can help to provide context for the content and improve its long-term preservation. Additionally, tracking the digital content throughout the preservation process is crucial to ensure that it is not lost or damaged. This process also enables archives to manage access and permissions for the content, safeguarding it against unauthorized access or modification. Overall, proper accessioning and tracking of digital content are fundamental components of any successful digital preservation initiative.

1.3.2 File integrity checking and fixity

File integrity checking is a crucial aspect of digital preservation, ensuring that digital files remain unaltered and uncorrupted over time. This process involves using hash functions like MD5 checksums to verify that the digital files have not been modified during transfer or storage. These functions generate a unique digital fingerprint for each file, and even a small change in the input results in a vastly different output, ensuring the authenticity and accuracy of the digital content (Garfinkel, 2013). Fixity checking ensures that digital content has remained unchanged over time. It guards against accidental or intentional modifications, hardware or software failures, or other factors that could result in data loss or corruption. By regularly checking the fixity of digital

content, archives can detect any unauthorized changes or tampering and take appropriate action to preserve the content's integrity.

In addition to file integrity checking, hash functions can be used for to identify duplicative digital content. If a “hash collision” occurs, where two files have identical checksums, the contents of the digital object should be examined for duplication (Garfinkel, 2013). Sloyan's (2016) cautionary statement regarding duplicate files highlights the importance of thoroughly examining the reasons for their existence before deciding to delete them. While it may seem logical to eliminate all duplicates to free up space and maintain a clean and organized system, there may be instances where duplicate files serve a unique purpose.

For example, in a folder of meeting minutes, there may be multiple duplicate files that serve as backups or different versions of the same document. In such cases, it would be unwise to delete these duplicates without first ensuring that the most recent version is saved and easily accessible. Similarly, in a project file, there may be duplicate files that uniquely support the association of those files, such as reference materials, images, or data sets that are used in multiple parts of the project. Deleting these duplicates without examining their purpose could disrupt the integrity of the project.

Sloyan also advises checking the last modification date stamp, which can provide valuable information about why the duplicates exist. If the modification dates are close together, it may suggest that the files were accidentally copied or saved multiple times. On the other hand, if the modification dates are far apart, it may indicate that the duplicates were intentionally created to serve different purposes.

Barrera-Gomez and Erway (2013) notes that while an altered checksum can alert a processing archivist to the fact that a file has been changed, it cannot indicate what the alteration was. This underscores the importance of maintaining multiple copies of digital content and regularly checking their fixity to detect and address any potential issues. Incorporating file integrity checking and fixity into the pre-ingest processing manual is critical for archives to ensure the accuracy and authenticity of their digital collections over time. By doing so, they can preserve the value of these collections as essential resources for research and other purposes.

1.3.3 Virus scanning and malware detection

Virus scanning and malware detection are critical to ensuring the security of digital content in archival collections. These processes are a key component of the pre-ingest processing manual, and they are essential to preventing potential threats to the archive's infrastructure. Virus scanning involves the use of antivirus software to scan digital content for known viruses and other malicious software that can harm the archive's system or compromise the security of the digital content. The importance of virus scanning and malware detection lies in the fact that malicious software can cause significant damage to digital content and the archive's infrastructure. Viruses can spread quickly through digital content and cause data loss, corruption, or even complete system failure. Malware can compromise the security of digital content by allowing unauthorized access to sensitive information, including personal information or confidential data.

Barrera-Gomez and Erway (2013) recommends that archives create a working copy of files from the master copy and then run virus scan software on the working copy. This ensures that the master copy and its associated metadata are not altered or damaged

during the virus scanning process. By incorporating virus scanning and malware detection as part of the pre-ingest processing manual, archives can protect their collections from harm and ensure the long-term accessibility and usability of their digital content. These processes are essential for maintaining the integrity and security of digital archives, and they should be regularly updated to stay current with new threats and emerging technologies.

1.3.4 File format assessment

File format assessment is a critical component of the pre-ingest processing manual for archival digital files. This process involves assessing the suitability of file formats for preservation, considering factors such as obsolescence and future accessibility. A 2016 survey conducted by DeRidder and Helms found that the top file formats were TIFF, WAV, PDF/A, MPEG-4, CSV, and TXT. These formats were identified as being the most suitable for long-term preservation because they are widely supported and have established standards that make them compatible with various software and hardware platforms.

DROID is a software tool used to identify file formats and is often used in pre-ingest processing of digital content. It uses PRONOM, a File Format Registry developed by the UK National Archives, to identify file types (“File profiling tool,” n.d.). PRONOM is a comprehensive database of technical information about file formats, including their extensions, magic numbers, and signatures. When DROID is employed to identify file types, it uses PRONOM's database to compare the unique characteristics of an unknown file against known file format signatures (Walsh, 2017). If a match is found, DROID can identify the file format, which is crucial information in the pre-ingest processing stage of

digital archiving. The ability to identify file formats using DROID and PRONOM provides archivists with critical information necessary to preserve digital content for long-term access and use.

File formats can become obsolete, rendering digital content inaccessible over time. Therefore, archives need to select file formats that are stable, widely used, and have long-term support. File format assessment involves examining the characteristics of file formats, such as their structural complexity, compression, and the presence of proprietary elements, which can affect long-term preservation. The choice of file format can significantly impact the longevity and accessibility of digital content, making it an essential component of the pre-ingest processing manual for archives.

1.3.5 Appraisal and selection

Effective appraisal and selection processes enable archives to build collections that reflect their missions and priorities, making them valuable resources for researchers, scholars, and the public. By preserving the most significant digital content, archives can offer insight and understanding of historical, cultural, and social contexts. In terms of digital content, Belovari (2019) distinguishes between *broad appraisal*, utilizing software to identify duplicates, junk files, and sensitive personal information and *qualitative appraisal* where an archivist considers the content, format, provenance, use, legal requirements, privacy, and access restrictions. As such, these archives play a vital role in the dissemination of knowledge and understanding to future generations.

In another article, Belovari (2017) details a workflow she used to process digital content at the German State Archives. Her process begins with a preliminary inspection, during which she asks questions, develops criteria, and identifies potential risks. From

there, she moves on to a broad appraisal, which involves both manual and software-based deduplication. In the manual deduplication phase, Belovari evaluates and deletes identical directories, while in the software deduplication phase, she reviews more granular duplicated files. Belovari also uses software to remove empty directories, as well as empty, temporary, and software files. Once the deduplication process is complete, Belovari appraises the remaining files qualitatively, using personal, content-related, and visual-related criteria. Personal criteria refer to the uniqueness of the material, while content-related criteria involve the historical or cultural significance of the material. Visual-related criteria, on the other hand, pertain to the format and appearance of the digital content. Overall, Belovari's process is comprehensive and systematic, starting with a quick inspection and progressing to a more detailed appraisal. By combining manual and software-based deduplication and using various criteria to appraise the remaining files, she can effectively manage and preserve digital collections.

It is crucial to have effective appraisal and selection processes because archives often have limited resources and must carefully allocate them to ensure they preserve the most valuable digital content. Niu (2014) suggests that a framework for appraisal should be developed based on an institution's selection criteria, legal issues, technical considerations, preservation factors, and the presence or absence of information and value judgments. When dealing with hybrid collections, those that contain a mixture of physical and digital materials, Belovari (2019) found that those digital carriers, disks and other removable media, do not have a consistent hierarchy or folder titles more typically found in large digital collections.

Belovari's (2017) research on the processing of materials highlights the challenges associated with managing digital collections. One of the main difficulties faced by archivists when processing digital materials is the sheer volume of data that needs to be appraised and organized. This can be a time-consuming task, particularly when dealing with large collections.

In addition to the time-consuming nature of the task, Belovari also observed that personal attitudes and biases can impact processing behaviors. For instance, when she became fatigued, she tended to discard more physical materials than digital ones. According to Belovari, "When I was tired, keeping digital files appeared to have little cost (handling, storage, etc.)" (p. 73). This may be due in part to the fact that digital files often appear as a simple list of file names on a screen, which can make it more difficult to evaluate their significance and value.

These observations highlight the importance of being mindful of personal attitudes and biases when processing digital collections. It's important for archivists to recognize that their perceptions of the relative value of physical versus digital materials may not always align with the actual value of those materials. Additionally, it's crucial to develop effective strategies for managing the large volume of digital data that needs to be appraised and organized.

To address these challenges, archivists can develop specialized tools and software for appraising and organizing digital files. Additionally, establishing guidelines and best practices for managing large volumes of digital data can help to ensure that important materials are properly preserved and accessible in the long-term. By remaining aware of personal attitudes and biases and developing effective strategies for managing digital

collections, archivists can help to ensure that these important materials are properly preserved and accessible to future generations.

1.3.6 Sensitivity Review

The pre-ingest processing manual for archival digital files includes a critical step known as sensitivity review. This is especially important for archives with sensitive or confidential materials. Sensitivity review involves assessing digital content to identify any potentially sensitive or confidential information that requires protection, such as personal identifying information (PII) and personal health information (PHI).

Additionally, sensitivity review determines appropriate access and use restrictions for sensitive materials, which are then recorded in the metadata. When there are thousands of digital items, it is improbable for a digital archivist to examine the content of each object. Sloyan (2016) conducted a case study on the hard drives at Wellcome Library and described a process of sampling files based on a categorical risk assessment of the collections. This assessment was made based on the available knowledge about the collection at the time of accession. The files that were considered to have the highest risk of containing sensitive information were reviewed. If any sensitive information was found in these files, access to the entire unit was restricted.

Tools such as BulkExtractor, a digital forensics software, can scan and flag files containing potentially sensitive information such as credit card numbers and social security numbers (Cirella, 2020). By incorporating sensitivity review into the pre-ingest processing manual, archives can safeguard sensitive digital content from unauthorized access or use, ensuring its security and accessibility only to authorized individuals. This

approach can foster trust and ensure compliance with relevant laws and regulations regarding the handling and use of sensitive information.

1.3.7 Metadata creation and management

Metadata creation and management are vital steps in the pre-ingest processing of archival digital files. Metadata refers to the descriptive information that describes the digital content, including its creator, subject, date, and other relevant details. The purpose of metadata is to provide context and enable the discovery and use of digital content by researchers and other users. Without proper metadata, digital content can be challenging to find and use effectively.

Creating metadata involves identifying and recording relevant information about digital content using standard formats and controlled vocabularies. It is essential to keep documentation of what was done to the content and who was involved in the metadata creation process. One approach to organizing metadata and other project-related documentation is to create a project directory that contains a master folder to hold the original copy of the content, a working folder for working copies, and a documentation folder for capturing metadata and other information (Barrera-Gomez & Erway, 2013). Metadata management involves ensuring that metadata is accurate, consistent, and up to date, as well as organizing and maintaining it in a structured manner. This involves regular updates and review of the metadata to ensure that it remains relevant and useful to users. By effectively managing metadata, digital content can be easily found, accessed, and used by researchers and other users, thereby enhancing its discoverability, accessibility, and usability.

1.4 Professional standards and guidelines for Pre-Ingest Processing

1.4.1 Open Archival Information System

The OAIS (Open Archival Information System) model is an abstract reference model that was designed to provide a framework for managing and preserving digital information. It was created to be a flexible framework that enables archival repositories to customize their implementation to meet the needs of specific user groups. The OAIS model defines the roles and responsibilities of different stakeholders involved in the digital preservation process, including producers, managers, and users of digital information (Consultative Committee for Space Data Systems, 2012).

One of the essential concepts introduced by the OAIS model is the archival package. An archival package comprises digital objects and their accompanying metadata, such as reference information, context information, provenance information, fixity information, and access information. Archival packages are stored in an archive and maintained in a manner that ensures their long-term preservation and accessibility. The OAIS model has become a common language that is widely understood by digital preservation practitioners from different professional backgrounds (Baucom, 2019). This has facilitated communication and collaboration between different stakeholders involved in digital preservation initiatives. The OAIS reference model has played a crucial role in advancing digital preservation practices and standards, providing a framework that guides best practices for managing and preserving digital information.

1.4.2 Digital Processing Workstation

Digital preservation workstations are essential tools for cultural heritage institutions and archives to effectively manage, process, and preserve their digital collections. Barrera-Gomez and Erway (2013) recommends that organizations keep a separate computer designated solely for processing and managing digital materials, which is not connected to any network. This approach reduces the risk of virus or malware exposure from unprocessed files, as well as unauthorized access to restricted materials. This non-networked computer can be used to transfer files or update software when necessary, and its sole purpose is to manage the organization's digital collections. By having a designated workstation for digital preservation activities, the risk of compromising the organization's network security is greatly reduced, and confidential or sensitive materials can be handled securely.

Assembling a digital preservation workstation is not a simple task, as noted by Arroyo-Ramirez, Bolding, Charlton, et al. (2018). It requires careful planning and consideration of the specific needs of the institution's digital collections. This process involves evaluating the hardware and software requirements of the workstation, selecting the appropriate software tools, and ensuring that the workstation is equipped with sufficient storage capacity and processing power to handle the institution's digital collections effectively. Princeton Rare Book Books and Special Collections staff recognized the need for a digital processing workstation to manage their born-digital content and gain intellectual and physical control of their collections. Initially, a dual booting Windows 7/BitCurator (Linux) portable laptop was acquired, but it was later replaced by a Forensic Recovery of Evidence Device (FRED) machine due to insufficient processing power and local storage.

Durno's (2016) case study on legacy floppy disks from the 1980s highlighted that using archival toolkits like BitCurator or Archivematica alone may not be sufficient for content recovery. Instead, a variety of specialized tools developed by the retro-computing and software preservation communities were found to be more effective in this scenario. It is important to recognize that digital preservation requires ongoing learning and adaptation to new technologies and tools to ensure long-term access to born-digital content.

1.5 Challenges of effective pre-ingest processing

According to Corrado (2022), there are several challenges associated with preserving digital content. One of the main challenges is the economic cost of preserving digital content, which includes the cost of acquiring and maintaining digital storage, hardware, software, and infrastructure. Additionally, skilled labor is required to manage and preserve digital content, which can be costly and challenging to obtain. Another challenge is the legal issues related to copyright, privacy, and ownership of digital content. Preserving digital content requires adherence to legal frameworks, which can be complex and require specialized knowledge. To overcome these challenges, organizations must implement effective preservation strategies that address economic, technical, and legal aspects of preserving digital content.

Johnston (2020) emphasizes that there is no perfect solution for digital preservation. Instead, she encourages each organization to find the approach that works best for them. As she notes, "There is no one best technology. There is no perfect workflow. There is no one right way. Do what makes sense for your organization. But you have to do something" (p. 197). This quote underscores the importance of taking

action when it comes to managing digital content. While it can be tempting to wait for the "perfect" technology or workflow to emerge, Johnston argues that such perfection is unattainable. Instead, organizations should focus on finding a system that works for them and taking steps to implement it. Of course, this is easier said than done. With so many options available, it can be difficult to know where to begin. Johnston's advice is to start small and build from there. This might mean identifying a single area of your digital content management that needs improvement and working on that first. As you gain experience and confidence, you can expand your efforts and refine your approach. Ultimately, the key takeaway from Johnston's quote is that action is essential when it comes to managing digital content. While there may not be a single "right" way to do it, doing something is better than doing nothing. By taking small steps and continuously improving, organizations can develop a digital content management strategy that works for them.

1.6 Conclusion

Pre-ingest processing is a critical stage in digital archives, ensuring the long-term preservation and accessibility of digital files. This review has highlighted the importance of pre-ingest processing in digital archives, as well as the different components that should be considered when developing a manual for pre-ingest processing. Future research and development in pre-ingest processing should focus on addressing the gaps and challenges identified in this review, such as the need for better tools and strategies for dealing with file format obsolescence, bit rot, and scalability. It is also important to acknowledge the community support that is essential in advancing digital preservation initiatives. Building on the work of the past is critical in digital preservation, and

collaboration and communication are essential for success (Baucom, 2019). Additionally, sustainability is a great concern for LIS practitioners, and frameworks and models need to be developed to improve processes of sustainability. The current state of pre-ingest processing in digital archives is evolving, and there is a need for ongoing research, collaboration, and development to ensure that digital archives continue to preserve and provide access to our cultural heritage for generations to come.

METHODOLOGY:

1.1 Positionality Statement:

Since October 2021, I have been interning at Duke University Medical Center Archive, working on a variety of projects related to hybrid collections. Throughout my time here, I have engaged with other staff members to gain insight into the current digital preservation workflow and Electronics Records Processing Guide. By doing so, I have identified the guide's deficiencies and worked closely with the archives team to understand their needs for an improved electronic records workflow and guide.

My approach to problem-solving is pragmatic and solution-focused, rather than being focused on understanding causation. This approach is heavily influenced by my undergraduate education in the works of early 20th-century Pragmatists at the Chicago School. My experience in a public high school library with limited resources further honed this approach, as I had to be strategic and creative to find solutions to problems.

When faced with limited resources, I believe it's essential to rely on practical and effective solutions in the moment. For instance, when a student showed signs of attention disorder, I provided effective strategies by breaking down large tasks into smaller ones and offering frequent feedback and redirection. Pragmatic thinking follows the evidence and takes the necessary steps to achieve the desired results. It is agile and adaptable to

new evidence, changing methods of intervention if the initial results don't meet the expected outcomes.

A pragmatic approach to digital preservation acknowledges the complex and evolving nature of technology and its impact on society. It emphasizes practical solutions that prioritize the preservation of digital materials for long-term access and use, while recognizing the limitations of resources, technology, and institutional support. This approach seeks to balance the competing priorities of accessibility, authenticity, and integrity of digital materials, while also considering the diverse needs and interests of stakeholders, including creators, users, and cultural heritage institutions. Ultimately, a pragmatic approach to digital preservation seeks to ensure that valuable digital materials are preserved for future generations in a sustainable and responsible manner.

1.2 Project Background

The need to revise Duke University Medical Center Archive's Electronic Records Processing Guide was realized over the summer months of 2022 when DUMCA received materials from the closing laboratory of retiring Duke University Professor, Dr. Paul L. Modrich. Modrich, a Nobel Prize recipient in 2015 for his contributions to the field of biochemistry, joined Duke University's faculty in 1976. The accession of nearly a half-century of mixed materials included two external computer hard drives. At that time, DUMCA had already a documented workflow for managing born-digital materials. The preexisting Electronic Records Processing Guide was developed several years earlier and was primarily designed for small-batch ingests of digital materials sourced from email attachments, cloud transfers, and other small-capacity digital carriers. It was a well-intended effort to initiate a digital archive, but the resulting workflow did not adhere to

the best practices as delineated by the OAIS reference model. For example, transferring electronic files from digital carriers did not utilize write blockers or hashing to preserve the integrity of these records. The lack of these protective measures allowed for inadvertent alterations to the electronic records during the process of capturing, and no means for determining if the captured copy of the digital object had been altered from its original state.

The receipt of not hundreds but thousands of electronic files became the impetus for a much-needed reevaluation of the workflow that had been sufficient for the short term but arguably not the best practice for long-term preservation and access to these materials. Also, the current procedure was completed by a labor-intensive review of each file by the accession's processing archivist. An improved workflow was needed that would employ software to report file metadata for appraisal and description, identify duplicated materials, and flag digital objects containing personal identifying information (PII).

In the months following the acquisition of Dr. Modrich's materials, the archive's staff and development team engaged in multiple conversations to determine the best approach for processing and preserving the digital content contained within. These discussions ultimately led to the decision to develop a comprehensive Pre-Ingest Processing Manual for Digital Content.

To ensure that the manual was effective and practical, the development and testing process initially focused on two smaller but still significant electronic accessions: the Garnett H. Kelsoe Laboratory Notebooks and the Physician Assistant Program Records. These accessions provided valuable insight into the unique challenges and considerations

associated with digitizing and preserving materials in the medical field. The Pre-Ingest Manual is designed to serve as a roadmap for the archive's team as they undertake the significant project of transferring, processing, and ingesting the digital content from the legacy carriers in Dr. Modrich's materials. Once the manual has been assessed and revised by the archives team, it is expected to play a vital role in ensuring the long-term preservation of these important materials for future generations.

1.3 Organization Description: Duke University Medical Center

Archives

Under the oversight of the Duke University School of Medicine and its parent organization Duke Health, the Medical Center Archives is a departmental unit within the Duke University Medical Center Library & Archives. The physical location of the archives is separate from the Duke Medical Center Library. Staff offices, processing work areas, storage facilities, and a reading room are in an off-campus warehouse building in northwest Durham, North Carolina. The archive also has an on-campus office at the Medical Center Library. With advance notice, this remote office at the Medical Center Library can be used as a space for patrons to meet with archives staff and access materials from the archives.

Although Duke University's Medical School and Hospital were founded in 1930, DUMCA was not established until 1977. In the years that followed a substantial number of records, photographs, publications, and recordings of interviews related to the history and business of the Medical Center were collected for an archive. Today, the archive predominantly serves the Duke University research community, although inquiries made by the public are welcome. Additionally, there is an online digital repository to aid

researchers with scanned historic photographs, publications, and exhibits related to Duke Health's history.

To carry out the work, the archive is composed of a team of four full-time and one part-time staff member to manage ten thousand linear feet of physical materials as well as born-digital files documenting Duke Health's history. Each member of the archive's staff is charged with unique responsibilities that are necessary to the department's function. The Director of the Medical Center Archives & Digital Library Initiatives, the Assistant Director for the Medical Center Archives and Head of Technical Services, and the Research, Outreach, and Education Librarian serve the essential core functions of the archives and are full time permanent salaried positions. The two remaining positions, Processing Archivist, and Intern are not permanent. They both provide support to the Head of Technical Services for ongoing archival projects. The Rice Diet Program Processing Archivist is a two-year contracted full-time position funded by a grant to prepare materials for researchers of the Rice Diet. Additionally, the archive offers a part-time internship to a graduate student of archival studies to process a variety of collections, and this is the position that the author of this paper currently holds.

1.4 AXAEM: An Integrated Records Management System

In 2017, the Medical Center Archive drafted the first version of its Electronic Records Processing Guide. The guide was jointly authored by the Head of Technical Services, Lucy Waldrop, and archive interns, Kahlee Leingang, and Alexandra Dowrey. Three considerable revisions of the word document have followed since the guide's original implementation. The guide provides the basic steps to prepare batches of electronic records for ingest into Axaem.

Axaem is an acronym for AppX-based Archives Enterprise Manager. The open-source application operates on an APPX platform, providing robust capabilities for physical and electronic records management. Axaem's Electronic Records Module is designed to comply with the OAIS reference model. The system provides several utilities for the creation and maintenance of the Archival Information Packages (AIP), and the generation of Dissemination Information Packages (DIP). When a Submission Information Package is ingested into Axaem from the Medical Center Archive's server, the system performs a chain of processes to ensure the electronic record ingest meets required OAIS benchmarks. SIPs are ingested into Axaem as a single or set of bags spawned according to BagIt specifications. Each bag contains the content and metadata that describes the contents transfer. Upon initiating the ingest, each electronic object in a bag is first linked to the associated bibliographic, transfer, and batch records. A virus scan is performed, and the bags are then again validated. Then metadata extractors including FITS, JHOVE2, Droid, and MediaInfo link descriptive information to each electronic object's associated record. Checksums are calculated for each electronic object and recorded in the database. After the ingest is complete, a report is generated which contains a list of the accepted records and those rejected due to any encountered errors.

To facilitate access and discovery of electronic records, during the ingest process, an access copy of each ingested object is created from the original. The access copy along with other information contained in the AIP is automatically supplied to the Dissemination Information Package (DIP). Axaem generates archival finding aids and catalog records and supports a searchable patron interface of indexed metadata. Of

course, access restrictions can be applied and are managed through the electronic object's bibliographic record.

1.5 An Important Consideration: HIPAA & FERPA

As an archive serving a private entity in the health and education industry, and as a subsidiary of Duke Health, the Duke Medical Center Archive is required by law to protect the privacy of individuals. The Family Education Rights and Privacy Act of 1974/1976 (FERPA), and the Health Insurance Portability and Accountability Act of 1996 (HIPAA) hold certain types of protected information that cannot be distributed or accessed without approved clearance. Although the Medical Center Archive is not responsible for active or inactive patient medical records and student records, it is still essential that members of the archives staff are vigilant for personal health information (PHI) and personal identifiable information (PII) when working with materials, because of the risk of legal penalties as an institutional covered entity. In some circumstances, some materials will be restricted by the archives under the law. Additionally, access to materials might be restricted because of university records policies or by requests made by the donor. Further explanation of specific access restrictions is published in each collection's online finding aid.

To ensure sensitive information is protected, it's important to take extra steps in the digital processing workflow, such as using software that performs cursory sensitivity reviews for Protected Health Information (PHI) and Personally Identifiable Information (PII). This software scans digital files to identify any sensitive information that requires further review or redaction.

By integrating this software into the digital processing workflow, archivists can ensure that sensitive information is properly handled, reducing the risk of unauthorized access or disclosure and ensuring compliance with relevant policies and regulations. Furthermore, the software can streamline the processing workflow by automating initial review and identification tasks, freeing up archivists' time to focus on more complex aspects of processing digital collections.

If the software flags any items, the processing archivist will review them. If a pattern of sensitivity emerges, the processing archivist will perform a systematic sensitivity review of all related materials. Although the software may not identify all sensitive information, it serves as a preliminary indicator that PHI or PII may exist in the collection, requiring additional steps before the collection can be made available.

1.6 Current Electronic Records Processing Guide Description

The current Electronic Records Processing Guide (June 2020) is a Microsoft Word document consisting of 50 pages of text and screenshot images. Since May 2017, the guide has been a working document that has been revised to accommodate procedural changes and information gaps. The guide is comprised of five sections. Section headings are interactively linked to a navigation pane containing an outline to facilitate easy recall of steps. Sections cover instructions for Electronic Records Processing Workflow, Bagger User Guide, Axaem's Electronic Records Processing Module, Running Reports, and Appendix.

The Electronic Records Processing Guide essentially supplies step-by-step instructions for the BagIt file packaging and ingesting of digital materials into Axaem's electronics records processing module which also can harvest a variety of descriptive and

technical metadata. The Electronic Records Processing Guide offers limited guidance for their migration from carriers, appraisal, description, and arrangement. DUMCA was also without a designated workstation and software to facilitate write blocking, disk imaging, and examination of the digital materials.

- ▼ Section 1: Electronic Records Processing Workflow
 - Section 1.1: Ingesting Records with an Accession Number
 - Section 1.2: Ingesting Records without an Accession Number
- ▼ Section 2: Bagger User Guide
 - Section 2.1: How to Download and Install Bagger to the Desktop
 - ▼ Section 2.2: Using Bagger
 - Section 2.2.1: Creating a New Bag
 - Section 2.2.2: Opening an Existing Bag
 - Section 2.2.3: Creating a Bag in Place
 - ▼ Section 2.3: Additional Features of Bagger
 - Section 2.3.1: Tag Files and Generated Metadata
- ▼ Section 3: AXAEM's Electronic Records Processing Module
 - Section 3.1: Ingesting Files into AXAEM
 - Section 3.2: Restricting Ingested Folders
 - Section 3.3: Restricting Ingested Files
 - ▼ Section 3.4: Deleting Ingested Records
 - Section 3.4.1: Deleting Entire Batch of Electronic Records
 - ▼ Section 3.4.2: Deleting an Individual File or Electronic Object
 - Section 3.4.2.1: Deleting an Individual File Record
 - Section 3.4.2.2: Deleting an Individual Electronic Object
- ▼ Section 4: Running Reports
 - Number and Size of Files Ingested
- ▼ Section 5: Appendix
 - Opening Bagger Using AXAEM
 - How to Search Accession Records on AXAEM

Figure 1 Screenshot of Navigation Pane from current Electronic Records Processing Guide (June 2020).

PROJECT OUTCOME:

1.1 Pre-Ingest Processing Manual for Digital Content

The result of this project is the expansion of Electronic Records Processing Guide by developing a Pre-Ingest Processing Manual for Digital Content, which serves as a document of record and means for conveying each step of digital preservation workflow to Medical Center Archives staff. In addition to the procedures already in place, this manual will include basic instructions for operating a designated desktop computer workstation loaded with a Linux operating system for digital forensics and digital preservation processing. Unlike individually assigned staff computer workstations, the shared Digital Preservation Workstation will have the advantage of being forensically sound, meaning it will disconnect from the network while users process born-digital materials to prevent unintended network interference and other storage networks from exposure to malware. The Digital Preservation Workstation will utilize the Linux distribution of Pop! OS. The choice to use a Linux operation system was the preference of the Director of Medical Center Archives and Digital Initiatives. In addition to these GUI applications, the Linux operating system utilities provide additional support through its command line interface. Instruction for basic command line functions will be included in the guide to support basic functions of anti-virus, deduplication, and sensitivity review applications using these command line interfaces.

The manual will be designed in such a way that it can be followed in logical order but also easy to consult as needed with sections identified for each task. The final length of the guide is undetermined. Its style will replicate the current guide so that it can be easily integrated. The Pre-Ingest Processing Manual for Digital Content will expand upon the previous version, adding

instructional sections on basic operation functions of the Digital Preservation Workstation using the Pop OS Linux distribution packages.

1.2 Introduction

Digital file processing is an essential step in preserving and providing access to archival materials in the digital age. The purpose of this manual is to provide guidelines and best practices for preparing digital files for ingest into our repository.

This manual covers a variety of topics related to pre-ingest digital file processing. We begin with Linux basics and helpful hints for managing digital files using command-line tools.

As part of our commitment to preserving the integrity of our digital materials, we provide guidance on virus scanning and fixity checks. Virus scanning helps to ensure that our files are not infected with malware or other harmful software, while fixity checks help to ensure that our files have not been altered or corrupted over time.

Next, we discuss the appraisal of digital files, including considerations for selecting and prioritizing files for ingest based on their historical value and potential for future research. We also cover deduplication, which is the process of identifying and removing duplicate files from our collections. This helps to reduce storage costs and streamline access to our digital materials. Finally, we discuss sensitivity review, which is the process of reviewing digital files to ensure that they do not contain sensitive or confidential information that may need to be redacted or restricted.

By following the procedures outlined in this manual, we can ensure that our digital files are properly processed, described, and stored for future generations of researchers and scholars. This manual is intended for staff who handle digital materials at

Duke University Medical Center Archives. Thank you for your commitment to preserving our collections and ensuring their accessibility for years to come.

In addition to the topics covered in this manual, documentation is an essential part of our pre-ingest digital file processing procedures. We provide guidance on creating and maintaining documentation related to our digital materials, including file inventories, preservation metadata, and access copies.

Effective documentation practices help to ensure that our digital files are properly managed and can be accessed and understood by future users. They also help to facilitate the ongoing preservation of our digital collections by providing a record of the actions taken to process and preserve our materials over time.

1.3 Linux Basics and Helpful Hints

If you're new to operating Linux Pop OS, here are some helpful hints to get you started.

- Mind your case: Linux commands are case-sensitive, so pay attention to uppercase and lowercase letters when typing commands.
- Stop command: To stop a command in the terminal, press the “**Ctrl+C**” keys.
- Clear the screen: Use the "**Clear**" command to wipe the terminal screen and make it easier to read.
- Print output: To print the output of a command into a text file, use the ">" symbol followed by the file name. For example, "ls > list.txt" will create a file called "list.txt" with the output of the "ls" command.
- Know your location: The "**pwd**" command will show you the address of the present working directory, so you always know where you are in the directory structure.

- Just use "**CD**" to return home: The "**CD**" command is a shortcut that returns you to the home directory.

1.4 System Update

To ensure a secure and efficient system operation, it's recommended to start each session with a package update by accessing the system's command line interface. To do this, you can follow these steps:

1. Open a Terminal window from the desktop taskbar.
2. Type the command "**sudo apt update**" and press enter.
3. The system will prompt you for your password. Type in your password (note: the cursor won't advance while typing your password), and press enter.
4. The system will display a report of available updates.
5. Enter the command "**sudo apt full-upgrade**".
6. The system will display a report of acquired updates and will prompt you with "Do you want to continue?". Type "**Y**" and press enter.
7. The system will display a report of completed updates.

1.5 Wired Connection (Disconnect/Connect)

These instructions describe how to connect and disconnect the Digital Preservation Workstation's wired connection. Wired connections are important for secure and efficient data transfer. It is important to disconnect when accessing or working with files that have not been scanned for viruses.

Disconnect Wired Connection:

1. Click the ethernet port icon in the top right corner of the screen on the desktop.



2. In the drop-down menu, select "Wired Connected" and then "Turn off."
3. This will disconnect the Digital Preservation Workstation from all connected networks, including the internet.

Connect Wired Connection:

1. Click the power icon in the top right corner of the screen on the desktop.



2. Click "Wired off" and select "Connect."
3. This will connect the Digital Preservation Workstation to the network, including the internet.

1.6 File Fixity

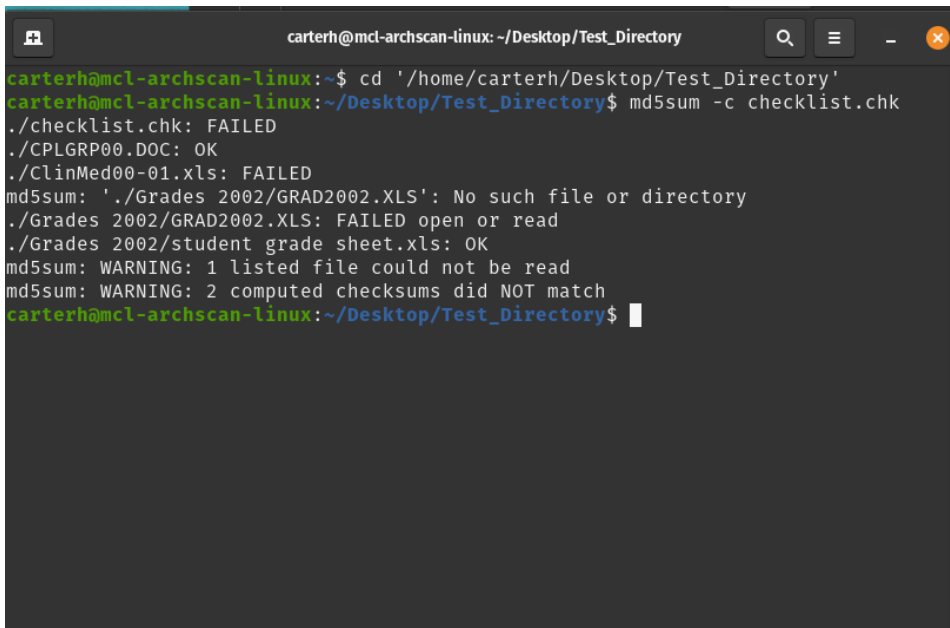
Checksums are important for ensuring file integrity and security. Checksums are unique values generated by applying an algorithm to a file. Comparing the checksum values of files with the values in a report can detect changes or tampering. It helps maintain data integrity, detect malicious activity, and ensure file security. Here are step-by-step instructions for generating a checksum report and employing a checksum check:

1.6.1 Generating a Checksum Report:

1. Open the terminal on your Linux system.
2. Navigate to the directory containing the files you want to generate checksums for.
3. Type the command "**md5sum * > md5sum.txt**" and press enter.
4. This command generates a list of checksums for any file in a specified directory and saves it to a file named "md5sum.txt".

1.6.2 Executing a Checksum Check:

1. Open the terminal on your Linux system.
2. Navigate to the directory containing the files you want to check.
3. Type the command "**md5sum -c md5sum.txt**" and press enter.
4. This command runs through the list of checksums in the "md5sum.txt" file to check them against the files in the directory.
5. If a file is missing or deleted, you can skip the warning prompt by adding "--ignore-missing" to the command: "**md5sum -c --ignore-missing md5sum.txt**".
6. At the end of the process, the system may output two types of warnings:
 - "md5sum: WARNING: # listed file could not be read", indicating that a file was likely deleted.
 - "md5sum: WARNING: # computed checksums did NOT match", indicating that a file has been altered either intentionally or unintentionally.



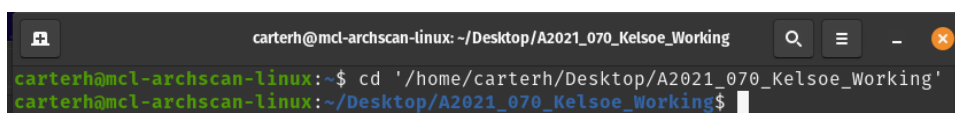
```
carterh@mcl-archscan-linux: ~/Desktop/Test_Directory
carterh@mcl-archscan-linux:~$ cd '/home/carterh/Desktop/Test_Directory'
carterh@mcl-archscan-linux:~/Desktop/Test_Directory$ md5sum -c checklist.chk
./checklist.chk: FAILED
./CPLGRP00.DOC: OK
./ClinMed00-01.xls: FAILED
md5sum: './Grades 2002/GRAD2002.XLS': No such file or directory
./Grades 2002/GRAD2002.XLS: FAILED open or read
./Grades 2002/student grade sheet.xls: OK
md5sum: WARNING: 1 listed file could not be read
md5sum: WARNING: 2 computed checksums did NOT match
carterh@mcl-archscan-linux:~/Desktop/Test_Directory$
```


1.7 Virus Scan using Clam AV

All transferred files should be scanned for virus before a Master File is uploaded to the E-Archives server and processing begins. Until files are scanned the digital preservation workstation's wired connection should remain turned off to prevent infected files from accessing Duke networks. Virus scans are initiated in the terminal using Clam AV. The software is installed to automatically update.

To access the software's help menu, enter the command **clamscan --help**

Once you have identified the files you wish to scan, open the terminal, using the "Change Directory" command **cd** to select the location of the folder containing the files.



```

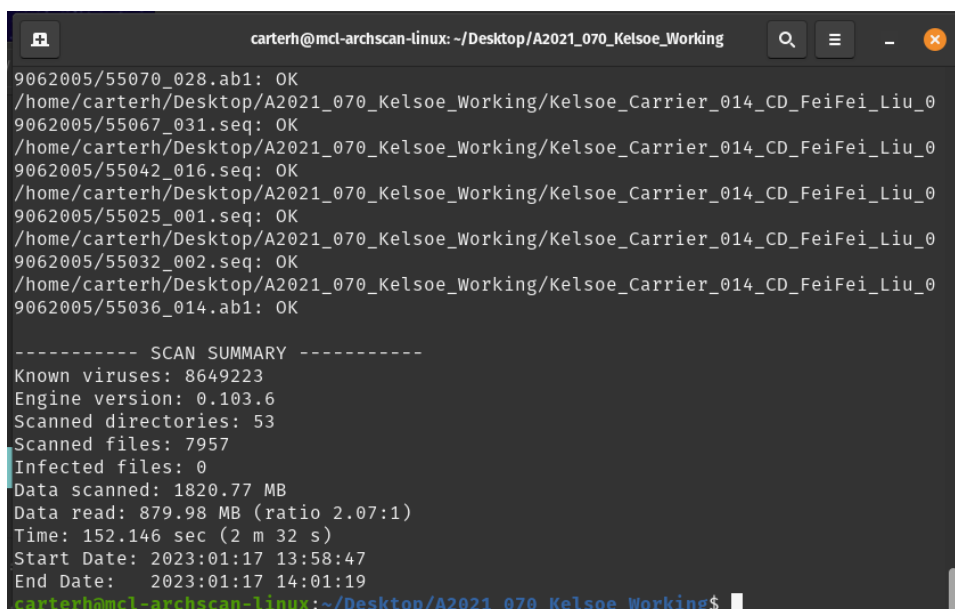
carterh@mcl-archscan-linux: ~/Desktop/A2021_070_Kelsoe_Working
carterh@mcl-archscan-linux:~$ cd '/home/carterh/Desktop/A2021_070_Kelsoe_Working'
carterh@mcl-archscan-linux:~/Desktop/A2021_070_Kelsoe_Working$

```

To scan all the contents of the folder and its subfolders.

Enter the command **clamscan --recursive**.

The virus scan will initiate. It may take some time for this process to occur. Once the scan is complete, a scan summary will print.



```

9062005/55070_028.ab1: OK
/home/carterh/Desktop/A2021_070_Kelsoe_Working/Kelsoe_Carrier_014_CD_FeiFei_Liu_0
9062005/55067_031.seq: OK
/home/carterh/Desktop/A2021_070_Kelsoe_Working/Kelsoe_Carrier_014_CD_FeiFei_Liu_0
9062005/55042_016.seq: OK
/home/carterh/Desktop/A2021_070_Kelsoe_Working/Kelsoe_Carrier_014_CD_FeiFei_Liu_0
9062005/55025_001.seq: OK
/home/carterh/Desktop/A2021_070_Kelsoe_Working/Kelsoe_Carrier_014_CD_FeiFei_Liu_0
9062005/55032_002.seq: OK
/home/carterh/Desktop/A2021_070_Kelsoe_Working/Kelsoe_Carrier_014_CD_FeiFei_Liu_0
9062005/55036_014.ab1: OK

----- SCAN SUMMARY -----
Known viruses: 8649223
Engine version: 0.103.6
Scanned directories: 53
Scanned files: 7957
Infected files: 0
Data scanned: 1820.77 MB
Data read: 879.98 MB (ratio 2.07:1)
Time: 152.146 sec (2 m 32 s)
Start Date: 2023:01:17 13:58:47
End Date: 2023:01:17 14:01:19
carterh@mcl-archscan-linux:~/Desktop/A2021_070_Kelsoe_Working$

```

If no viruses are detected, then the files can continue to be processed.

If a virus is detected, consult with the Director of the Archives to discuss options.

1.8 Creating Copies, Master Files, Ingest Files, and Working Files

After performing a virus scan, multiple copies of the files should be created so that there are three duplicated sets of files. (Think LOCKSS! ... Lots of Copies Keeps Stuff Safe!)

1.8.1 Multiple copies, their purpose, and where to store them?

Master Copy: These files should be moved to the E-Archives server where it serves as a backup to the Working Copy of files. This folder should not be accessed or modified throughout the processing project, unless there has been a loss or corruption of the Production File or Working Copy File.

Working Copy: This copy of files is created to facilitate processing decisions. Use these files to review their contents. These can remain on the desktop of the Digital Preservation Workstation until the processing project is complete.

Ingest Copy: This is the copy of files that will be used to ingested into Axaem, and should reflect the final selection of materials for long term preservation. These files remain on the desktop of the Digital Preservation Workstation until the processing project is complete.

1.9 Reviewing Files

What should we keep? Generally, we keep files that created by humans. Most files fall within this category. However, there are also files which are generated files by computers which can be easily identified and deleted. These include the files within RESOURCE.FRK and FINDER.DAT folders. Files which precede with a tilde character

(~) are also a good candidate for weeding as they are created by software to be used to rescue the document you are working on (all in theory). For example, `advance_article.doc` would be created as `~vance_article.doc`.

A helpful resource for identifying and describing digital file formats is the [PRONOM Technical Registry](#).

If uncertain, it's always okay to ask for help.

1.9.1 File Type Identification and Count

The below command lists and counts all file extensions recursively within the current directory.

`find . -type f | sed -n 's/.*\./p' | sort | uniq -c`



```

carterh@mcl-archscan-linux: ~/Desktop/A2021_070_Kelsoe_Ingest
carterh@mcl-archscan-linux:~/Desktop/A2021_070_Kelsoe_Ingest$ find . -type f | sed -n 's/.*\./p' | sort | uniq -c
4180 ab1
 18 bip
   1 doc
   2 exe
   1 EXE
   1 L01
   1 L02
   1 L03
   1 L04
   1 L05
   1 L06
   1 L07
   1 L08
   1 L09
   1 L10
   3 pdb
   2 pri
   1 psd
4094 seq
  36 Seq
   5 tif
   1 txt
  14 xls
   1 xlsx

```

Note the source for this output is the file name. File names may not resemble the actual file property. For a more consistent output, use this program. (Note: This program has to be initiated within the home directory).

`cd -`

`bash ./filestats.sh /directoryaddress/`

Directory Address Example: `/home/carterh/Desktop/A2022_012_Ingest_Work`

```

carterh@mcl-archscan-linux:~/Desktop/A2022_012_Ingest_Work$ cd -
/home/carterh
carterh@mcl-archscan-linux:~$ bash ./filestats.sh /home/carterh/Desktop/A2022_012_Ingest_Work
===== File types and counts =====
JPEG image data : 2481
Microsoft Access Database : 6
DOS EPS Binary File Postscript starts at byte 27610 length 103649 TIFF starts at byte 30 length 27580 : 1
Composite Document File V2 Document : 4392
CDFV2 Microsoft Excel : 3
PDF document : 462
TIFF image data : 794
RIFF (little-endian) data : 5
WordPerfect document : 4
OpenPGP Secret Key : 2
DOS EPS Binary File Postscript starts at byte 30 length 302435 TIFF starts at byte 302465 length 9733 : 1
VAX-order2 68k Blit mpX/muX executable : 1
CSV text : 23
MIPSEL-BE MIPS-III ECOFF executable not stripped - version 0.0 : 5
PGP Secret Sub-key - : 1
locale data table : 1
HTML document : 99
Microsoft ASF : 1
OpenDocument Spreadsheet : 1
Sony PlayStation PSX image : 1
VAX-order 68k Blit mpX/muX executable : 2
PC bitmap : 28
GIF image data : 343
PNG image data : 1
Adobe Photoshop Image : 11
OpenPGP Public Key : 2
ERROR: Output buffer space exceeded 1032+93 : 1
MPEG ADTS : 44
Non-ISO extended-ASCII text : 37
Microsoft Word 2.0 Document : 2
PE32 executable (GUI) Intel 80386 : 1
ASCII text : 186
GLF_BINARY_LSB_FIRST : 1
ISO-8859 text : 2
DOS EPS Binary File Postscript starts at byte 30 length 357510 TIFF starts at byte 357540 length 128482 : 1
data : 909
Rich Text Format data : 35
carterh@mcl-archscan-linux:~$ vi filestats.sh
carterh@mcl-archscan-linux:~$ bash ./filestats.sh /home/carterh/Desktop/A2022_012_Ingest_Work > carter.txt

```

1.9.2 List All Files Ordered by Size

The below command lists all files recursively within the directory by size (largest to smallest). This is one way to check for zero-byte files. These files can be deleted since they do not contain any data.

ls -laShR

```

carterh@mcl-archscan-linux:~/Desktop/A2021_070_Kelsoe_Ingest$ ls -laShR
.:
total 416K
drwxrwxrwx 2 carterh carterh 36K Jan 13 09:34 Nelson_Carrier_015_CD_Fe1Fe1_L10_PC_DNA_Sequences_11302000
drwxrwxrwx 2 carterh carterh 20K Jan 11 12:36 Nelson_Carrier_015_CD_Fe1Fe1_L10_PC_DNA_Sequences_11302000
drwxrwxrwx 2 carterh carterh 20K Jan 11 14:46 Nelson_Carrier_022_CD_Fe1Fe1_L10_PC_DNA_Sequences_11302000
drwxrwxrwx 2 carterh carterh 20K Jan 11 16:10 Nelson_Carrier_028_CD_Fe1Fe1_L10_PC_DNA_Sequences_11302000
drwxrwxrwx 2 carterh carterh 20K Jan 13 08:36 Nelson_Carrier_029_CD_Fe1Fe1_L10_PC_DNA_Sequences_11302000
drwxrwxrwx 2 carterh carterh 16K Jan 11 10:49 Nelson_Carrier_086_CD_Fe1Fe1_L10_PC_DNA_Sequences_11302000
drwxrwxrwx 2 carterh carterh 16K Jan 11 11:13 Nelson_Carrier_087_CD_Fe1Fe1_L10_PC_DNA_Sequences_11302000
drwxrwxrwx 2 carterh carterh 16K Jan 11 13:48 Nelson_Carrier_089_CD_Fe1Fe1_L10_PC_DNA_Sequences_11302000
drwxrwxrwx 2 carterh carterh 12K Jan 11 10:28 Nelson_Carrier_089_CD_Fe1Fe1_L10_PC_DNA_Sequences_11302000
drwxrwxrwx 2 carterh carterh 12K Jan 11 11:31 Nelson_Carrier_089_CD_Fe1Fe1_L10_PC_DNA_Sequences_11302000
drwxrwxrwx 2 carterh carterh 12K Jan 11 12:09 Nelson_Carrier_012_CD_Fe1Fe1_L10_PC_DNA_Sequences_11302000
drwxrwxrwx 2 carterh carterh 12K Jan 11 12:47 Nelson_Carrier_016_CD_Fe1Fe1_L10_PC_DNA_Sequences_11302000
drwxrwxrwx 2 carterh carterh 12K Jan 11 13:05 Nelson_Carrier_017_CD_Fe1Fe1_L10_PC_DNA_Sequences_11302000
drwxrwxrwx 2 carterh carterh 12K Jan 11 13:24 Nelson_Carrier_018_CD_Fe1Fe1_L10_PC_DNA_Sequences_11302000
drwxrwxrwx 2 carterh carterh 12K Jan 11 14:01 Nelson_Carrier_026_CD_Fe1Fe1_L10_PC_DNA_Sequences_11302000
drwxrwxrwx 2 carterh carterh 12K Jan 11 14:14 Nelson_Carrier_021_CD_Fe1Fe1_L10_PC_DNA_Sequences_11302000
drwxrwxrwx 2 carterh carterh 12K Jan 11 14:57 Nelson_Carrier_025_CD_Fe1Fe1_L10_PC_DNA_Sequences_11302000
drwxrwxrwx 2 carterh carterh 12K Jan 11 15:11 Nelson_Carrier_026_CD_Fe1Fe1_L10_PC_DNA_Sequences_11302000
drwxrwxrwx 2 carterh carterh 12K Jan 11 15:21 Nelson_Carrier_026_CD_Fe1Fe1_L10_PC_DNA_Sequences_11302000
drwxrwxrwx 2 carterh carterh 12K Jan 11 15:39 Nelson_Carrier_026_CD_Fe1Fe1_L10_PC_DNA_Sequences_11302000
drwxrwxrwx 2 carterh carterh 12K Jan 11 15:49 Nelson_Carrier_027_CD_Fe1Fe1_L10_PC_DNA_Sequences_11302000
drwxrwxrwx 2 carterh carterh 12K Jan 13 08:50 Nelson_Carrier_031_CD_Fe1Fe1_L10_PC_DNA_Sequences_11302000
drwxrwxrwx 2 carterh carterh 12K Jan 13 09:46 Nelson_Carrier_033_CD_Fe1Fe1_L10_PC_DNA_Sequences_11302000

```

This is one method to check for zero-byte files. These files can be deleted since they do not contain any data.

1.9.3 List All Files by Last Modified Date

This command will list recursively all files within a specified directory by their modification time (newest first). This will help identify a descriptive date range for the files, as well as show any inadvertent modifications made to the directory and its contents.

ls -ltR

```

carterh@mcl-archscan-linux:~/Desktop/A2021_070_Kelsoe_Ingest$ ls -ltR
.:
total 396
drwxrwxrwx 2 carterh carterh 4096 Jan 13 10:00 Kelsoe_Carrier_017_Laurens_DNA_Rodeo
drwxrwxrwx 2 carterh carterh 4096 Jan 13 09:59 Kelsoe_Carrier_025_CD_Feifei_Liu_PC
drwxrwxrwx 2 carterh carterh 4096 Jan 13 09:53 Kelsoe_Carrier_030_CD_Feifei_Liu_07052000
drwxrwxrwx 2 carterh carterh 12288 Jan 13 09:46 Kelsoe_Carrier_032_CD_Feifei_Liu_07012000
drwxrwxrwx 2 carterh carterh 36864 Jan 13 09:34 Kelsoe_Carrier_032_CD_Feifei_Liu_06202000
drwxrwxrwx 2 carterh carterh 12288 Jan 13 08:50 Kelsoe_Carrier_031_CD_Feifei_Liu_06272000
drwxrwxrwx 2 carterh carterh 4096 Jan 13 08:39 Kelsoe_Carrier_030_CD_Feifei_Liu_06122000
drwxrwxrwx 2 carterh carterh 20480 Jan 13 08:36 Kelsoe_Carrier_029_CD_Feifei_Liu_06002000
drwxrwxrwx 2 carterh carterh 20480 Jan 11 16:10 Kelsoe_Carrier_028_CD_Feifei_Liu_06002000
drwxrwxrwx 2 carterh carterh 12288 Jan 11 15:49 Kelsoe_Carrier_027_CD_Feifei_Liu_PC_DNA_S8

```

1.9.4 Identifying and Removing Zero-Byte Files?

Since files having zero bytes have no data, they can be removed. To identify zero-byte files recursively within a directory use the command:

find -size 0c

Alternatively, **find -type f -empty** can be used.

```

Workspaces Applications Mar 3 1:20 PM
carterh@mcl-archscan-linux:~/Desktop/A2022_012_Ingest_Work
carterh@mcl-archscan-linux:~/Desktop/A2022_012_Ingest_Work$ find -size 0c
./digital_files/CD_02/PAP240/Admin/fsaB.tmp
./digital_files/DVD_01/Photos 2003/PA Day 2003/SIV49.tmp
./digital_files/DVD_01/Class of 2005/Halloween 2003/Halloween/TMP9.tmp
./digital_files/DVD_01/Class of 2005/Halloween 2003/Halloween/halloweenie.jpg
./digital_files/DVD_01/Class of 2005/Halloween 2003/Halloween/TMP8.tmp
./digital_files/DVD_01/Class of 2005/Halloween 2003/Halloween/halloween0012.jpg
./digital_files/DVD_01/Class of 2005/Halloween 2003/Halloween/halloween1.jpg
./digital_files/DVD_01/Class of 2005/Halloween 2003/Halloween/TMP7.tmp
./digital_files/DVD_01/Class of 2005/Vegas Baby/SIV7A.tmp
./digital_files/DVD_01/Class of 2005/Vegas Baby/SIV139.tmp
./digital_files/DVD_01/Class of 2005/Vegas Baby/SIV4.tmp
./digital_files/DVD_01/Class of 2005/Vegas Baby/SIV42.tmp

```

To remove all zero-byte files recursively within a directory, use the command.

find -size 0c -delete

Alternatively, **find -type f -empty -delete** can be used.

It's always better to check the find results before using the -delete option.

To confirm that files have been deleted, repeat the **find -size 0c** command.

Identifying and Removing Empty Directories

To identify recursively empty directories (folders containing no items) use the command within a specified directory:

find -type d -empty

After reviewing the output directories can be deleted manually or by a batch. Enter the following command to delete all empty directories within a specified directory.

find -type d -empty -delete

To confirm that files have been deleted, repeat with the **find -type d -empty** command.

1.9.5 Counting Files

To count the number of files that occur recursively, use the following commands:

find <directoryaddress> -type f | wc -l

Or

tree <directoryaddress>

1.9.6 Calculating directory size (bytes)

To get total size of all the files in and under this directory

du -s (output in bytes)

Du -sh (readable expression such as MB, GB, etc.)

1.10 What about duplicates? Using fdupes

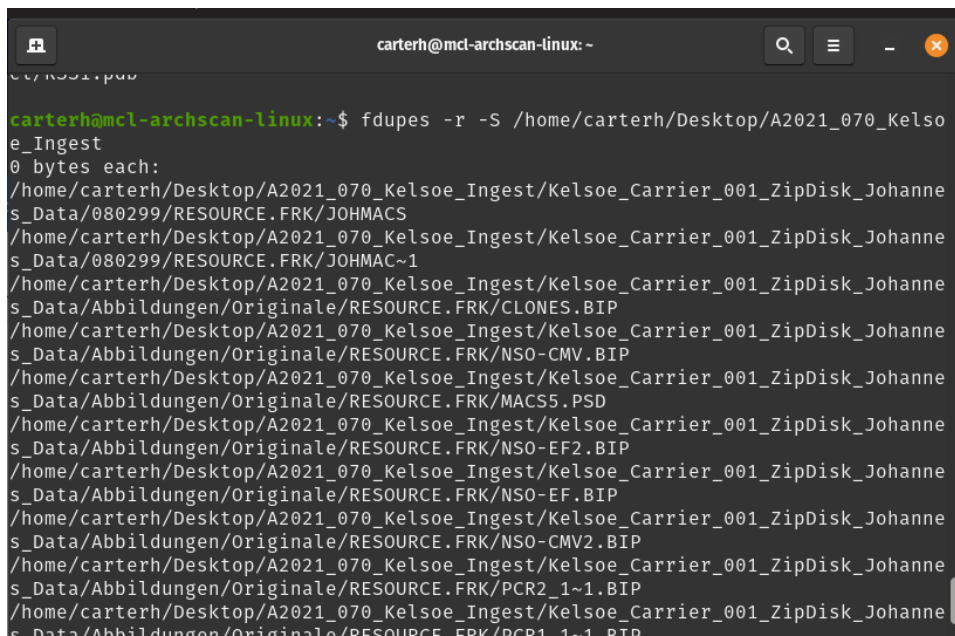
Generally, we want to avoid keeping duplicated files whenever possible. One software tool that can be used to identify the occurrence of duplicate files is `fdupes`. `Fdupes` identifies potential duplicates by comparing the checksums values within a specified directory. Just because items are flagged doesn't guarantee the files are duplicates. A file should always be reviewed before a decision is made to remove the file.

The program is initiated in the terminal using `fdupes`. The software is installed to automatically update.

To access the software's help menu, enter the command `fdupes --help`

Once you have identified the files you wish to scan, open the terminal, then use using `fdupes -r -S` followed by the directory's address.

For example: `fdupes -r -S /home/carterh/Desktop/`



```

carterh@mcl-archscan-linux: ~
carterh@mcl-archscan-linux:~$ fdupes -r -S /home/carterh/Desktop/A2021_070_Kelsoe_Ingest
0 bytes each:
/home/carterh/Desktop/A2021_070_Kelsoe_Ingest/Kelsoe_Carrier_001_ZipDisk_Johannes_Data/080299/RESOURCE.FRK/JOHMACS
/home/carterh/Desktop/A2021_070_Kelsoe_Ingest/Kelsoe_Carrier_001_ZipDisk_Johannes_Data/080299/RESOURCE.FRK/JOHMAC~1
/home/carterh/Desktop/A2021_070_Kelsoe_Ingest/Kelsoe_Carrier_001_ZipDisk_Johannes_Data/Abbildungen/Originale/RESOURCE.FRK/CLONES.BIP
/home/carterh/Desktop/A2021_070_Kelsoe_Ingest/Kelsoe_Carrier_001_ZipDisk_Johannes_Data/Abbildungen/Originale/RESOURCE.FRK/NSO-CMV.BIP
/home/carterh/Desktop/A2021_070_Kelsoe_Ingest/Kelsoe_Carrier_001_ZipDisk_Johannes_Data/Abbildungen/Originale/RESOURCE.FRK/MACS5.PSD
/home/carterh/Desktop/A2021_070_Kelsoe_Ingest/Kelsoe_Carrier_001_ZipDisk_Johannes_Data/Abbildungen/Originale/RESOURCE.FRK/NSO-EF2.BIP
/home/carterh/Desktop/A2021_070_Kelsoe_Ingest/Kelsoe_Carrier_001_ZipDisk_Johannes_Data/Abbildungen/Originale/RESOURCE.FRK/NSO-EF.BIP
/home/carterh/Desktop/A2021_070_Kelsoe_Ingest/Kelsoe_Carrier_001_ZipDisk_Johannes_Data/Abbildungen/Originale/RESOURCE.FRK/NSO-CMV2.BIP
/home/carterh/Desktop/A2021_070_Kelsoe_Ingest/Kelsoe_Carrier_001_ZipDisk_Johannes_Data/Abbildungen/Originale/RESOURCE.FRK/PCR2_1~1.BIP
/home/carterh/Desktop/A2021_070_Kelsoe_Ingest/Kelsoe_Carrier_001_ZipDisk_Johannes_Data/Abbildungen/Originale/RESOURCE.FRK/PCR1_1~1.BIP

```

This generates a list of potential files which should be reviewed for duplication.

1.11 What about PII? Sensitivity Review using Bulk Extractor and Risk Assessment

As a protected health entity, we have a legal obligation to prevent the dissemination of protected health information and personal identifiable information. Ideally, the only way to be completely confident that no information is ever shared with unauthorized individuals requires a complete review of all digital. Realistically, this is impractical given the volume of these materials and our limited resources. The best method to date of conducting a sensitivity review involves adequate understanding of the materials' provenance (do we know if the creators utilized PHI/PII?), sharp observation (where have I seen PHI/PII occur in the materials?), and sound assessment of the risk of PHI & PII (have we made genuine and deliberate effort to identify PHI/PII?) occurring within the collection of digital objects. One tool which may be helpful is Bulk Extractor the software is capable (but not 100% accurate) of identifying unique information within a directory of files. Such categories of information which can be identified include, such as phone numbers and credit card numbers and social security numbers. More specific information about Bulk_Extractor can be found [here](#).

To initiate a BulkExtractor report within the current directory and subdirectories use the following command.

```
bulk_extractor -R -o output /directory address/
```

For example: **bulk_extractor -R -o output**

```
/home/carterh/Desktop/A2021_070_Kelsoe_Working
```



```

carterh@acl-archscan-linux:~/Desktop/output$ bulk_extractor -R -o output /home/carterh/Desktop/A2021_070_
Kelsoe_Working cd /home/carterh/Desktop
mkdir "output"
bulk_extractor version: 2.0.1
Input file: "/home/carterh/Desktop/A2021_070_Kelsoe_Working"
Output directory: "output"
Disk Size: 8373
Scanners: aes base64 elf evtX exif facebook find gzip httplogs json kml_carved msxml net ntfsindx ntfslog
file ntfsmft ntfsusn pdf rar sqlite utmp vcard_carved windirs winlnk winpe winprefetch zip accts email gp
s
Threads: 8
going multi-threaded...( 8 )
bulk_extractor      Tue Jan 31 14:10:33 2023

available_memory: 26941964288
bytes_queued: 0
depth0_bytes_queued: 0
depth0_sbufs_queued: 0
elapsed_time: 0:00:00
estimated_date_completion: 2023-01-31 14:10:32
estimated_time_remaining: n/a
fraction_read: 0.000000 %
max_offset: 0
sbufs_created: 0
sbufs_queued: 0
sbufs_remaining: 0
tasks_queued: 0

```

After the process is completed the location of the output folder will be in the directory of which the command was initiated.

To view the resulting instances where specified strings were identified by the software.

Use “Change Directory” command **cd** to select the location of the folder containing the files. Then enter **ls -s** to list the files by size.

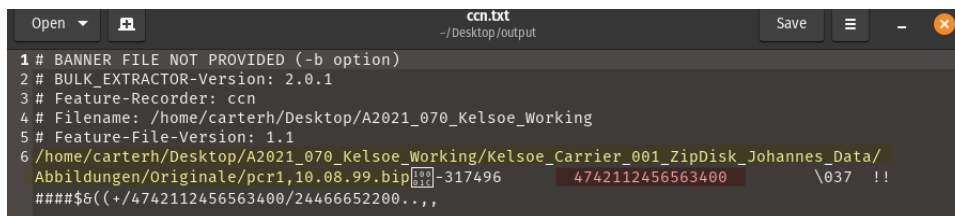
```

carterh@acl-archscan-linux:~/Desktop/output$ ls -s
total 2092
 0 aes_keys.txt           0 httplogs.txt          0 telephone.txt
 0 alerts.txt            0 ip_histogram.txt     0 unrar_carved.txt
 4 ccn_histogram.txt    0 ip.txt               0 url_facebook-address.txt
 0 ccn_track2_histogram.txt 4 jpeg_carved         0 url_facebook-id.txt
 0 ccn_track2.txt       4 jpeg_carved.txt     4 url_histogram.txt
 4 ccn.txt              24 json.txt            0 url_microsoft-live.txt
 4 domain_histogram.txt 0 kml_carved.txt      0 url_searches.txt
 92 domain.txt          0 ntfsindx_carved.txt 4 url_services.txt
 0 elf.txt              0 ntfslogfile_carved.txt 12 url.txt
 4 email_domain_histogram.txt 0 ntfsmft_carved.txt 4 utmp_carved
 4 email_histogram.txt  0 ntfsusn_carved.txt 16 utmp_carved.txt
24 email.txt           0 pii_teamviewer.txt  0 vcard.txt
 0 ether_histogram_1.txt 0 pii.txt              4 windirs.txt
 0 ether_histogram.txt  0 rar.txt              0 winlnk.txt
 0 ether.txt            1844 report.xml         4 winpe_carved
 0 evtX_carved.txt      0 rfc822.txt           4 winpe_carved.txt
 8 exif.txt             0 sin.txt              4 winpe.txt
 0 facebook.txt         0 sqlite_carved.txt   0 winprefetch.txt
 0 find_histogram.txt   0 tcp_histogram.txt    4 zip
 0 find.txt             0 tcp.txt              12 zip.txt
 0 gps.txt              0 telephone_histogram.txt

```

The numbers next to the file names indicate the file size and show that several of the files, including email.txt and domain.txt, were populated with features during the run.

Each of the report files provide a listing of flagged results which are identified by file address and a portion of the relevant content. These files should likely undergo a more intensive screening for sensitive information.



```

1 # BANNER FILE NOT PROVIDED (-b option)
2 # BULK_EXTRACTOR-Version: 2.0.1
3 # Feature-Recorder: ccn
4 # Filename: /home/carterh/Desktop/A2021_070_Kelsoe_Working
5 # Feature-File-Version: 1.1
6 /home/carterh/Desktop/A2021_070_Kelsoe_Working/Kelsoe_Carrier_001_ZipDisk_Johannes_Data/
  Abbildungen/Originale/pcr1,10.08.99.bip[REDACTED]-317496      4742112456563400      \037  !!
#####$(+/4742112456563400/24466652200.,.,

```

In the example above, a number resembling a credit card number appeared in the file, but after screening the digital object, it was apparent that this string of 16 digits was not a credit card number. (also, the file type can be a clue).

Special attention should be given to ccn.txt (Credit Card #s), and pii.txt (Social Security #s) files.

If sensitive information is observed in a digital object. The Head of Technical Services should be notified, as well as a plan for separation these digital objects be developed.

Segmentation fault (core dumped) Error Message – If this message is encountered the size of the directory you are attempting to scan is too large. You may need to scan a directory of files in smaller sets.

1.12 Renaming Files

Legacy file names may contain special characters, which can be the cause for ingest issues with Axaem. Using the software Bulk Rename Utility, file and folder names containing special characters can be identified and renamed as an aggregate.

A program file (.bru) has been created to aid in normalizing files names. The rules for the program replicate archival practices at other institutions.

- Replace spaces with underscores.
- Remove occurrences of double spaces and replace them with single spaces.
- Replace “&” with “and”
- Remove special characters including ~`!@#\$%^*()+={}[];";'<>? - +={}[];";'<>?/\.
- Remove accented characters and replace them with non-accented versions.
- Trim leading and trailing spaces. Never change the file extension through renaming. Changing a file extension can create errors that can make the file unreadable and essentially lost.

ASSESSMENT

The assessment plan aims to evaluate the effectiveness of the Pre-Ingest Processing Manual for Digital Content, which has been completed as a first draft. The plan outlines the steps that will be taken to assess the guide's usability and to make revisions as necessary.

The first step of the assessment plan involves scheduling work sessions with the Director and Assistant Director for the Medical Center Archives. These sessions will be conducted individually to ensure that each participant has an opportunity to provide their feedback and ask questions as they process and ingest digital materials using the guide. During these work sessions, the observer will document the participant's progress and note any questions or feedback that they provide. This information will be collected and reviewed to identify areas where the guide may be improved or clarified. Based on the feedback received, changes will be made to the Pre-Ingest Processing Manual for Digital Content and the Director and Assistant Director will be asked to review these revisions. If they are not satisfied with the resulting changes, additional revisions will be made until the guide is deemed acceptable.

This assessment plan provides a clear process for evaluating the effectiveness of the Electronic Records Processing Guide and ensuring that it meets the needs of the Medical Center Archives. By gathering feedback from key stakeholders and making

revisions as necessary, the guide will be more effective and useful for processing and ingesting digital material

DISCUSSION:

The development of a Pre-Ingest Processing Manual for Digital Content was a complex and challenging process that required careful planning, collaboration, and attention to detail. This manual is intended to serve as a comprehensive guide for the transfer, processing, and ingestion of digital materials into the archive's digital repository. In this section, I discuss the process that was undertaken to develop this manual, including the key stakeholders involved, the steps taken to create the manual, and the challenges that were encountered along the way.

1.1 Key Stakeholders

The development of the Pre-Ingest Processing Manual for Digital Content was a collaborative effort involving multiple stakeholders within the archive's organization. These stakeholders included the archive's staff and development team, the Director and Assistant Director for the Medical Center Archives, and other subject matter experts who provided input and guidance throughout the process. The involvement of these stakeholders was critical in ensuring that the manual was developed in a way that met the needs of the archive and its users. By working collaboratively, the stakeholders were able to provide valuable input and feedback at every stage of the process, ensuring that the manual was accurate, comprehensive, and effective.

1.2 Steps Taken

The first step was planning and scoping, which involved identifying the manual's objectives, defining its scope and purpose, and identifying the key stakeholders involved in the process. This involved outlining the specific goals and outcomes that the manual should address, such as streamlining the processing and ingesting of digital materials and ensuring the long-term preservation of digital materials. Defining the manual's scope and purpose was also a crucial aspect of this phase. I needed to establish the scope of the manual in terms of the type of digital materials it would cover, such as born-digital records, digitized records, or web archives. Additionally, I identified the manual's purpose, which was to provide a clear, concise, and user-friendly guide for the archive's staff to follow when processing and ingesting digital materials.

The next step was research and analysis, where existing literature and best practices were reviewed to determine the specific requirements and challenges associated with processing digital materials in the archive's context. I analyzed the available tools, techniques, and technologies required to develop a comprehensive manual that would meet the archive's needs. This step provided the team with a deep understanding of the subject matter and ensured that the manual's content was accurate and up to date. Based on the research and analysis, I drafted the manual and tested it in a controlled environment. I worked with smaller electronic accessions, such as the Garnett H. Kelsoe Laboratory Notebooks and the Physician Assistant Program Records, to identify any issues or challenges with the manual. This step helped me to identify gaps in the manual's content and refine it to make it more effective.

After testing the manual, feedback is collected and used to revise and review the manual. This ensures that the manual is accurate, comprehensive, and effective. The revision and review process is crucial in refining and improving the manual based on feedback from key stakeholders. Once the manual has been revised and reviewed, it is finalized and integrated into the archive's workflows and procedures for processing and ingesting digital materials.

To ensure that the manual is suitable for practical use, it is essential to implement and test it in a real-world context. This testing process helps to verify that the manual can be applied effectively and achieves its intended outcomes. This helps to ensure that archivists can use the manual effectively to guide their work in processing and ingesting digital materials. By integrating the manual into the archive's workflows and procedures, the archive can maintain consistent practices and ensure that materials are processed and ingested in a way that meets the needs of the archive and its stakeholders.

1.3 Challenges Encountered

Despite careful planning and collaboration, the development of the Pre-Ingest Processing Manual for Digital Content was not without its challenges. As part of the development process for the Pre-Ingest Processing Manual for Digital Content, I had to learn how to use Linux, a powerful and widely used operating system that is commonly employed in digital archiving and other technical fields. At first, I was quite apprehensive about learning this new system, particularly the command line interface that can be intimidating for beginners. However, as I began to work with Linux more and more, I began to appreciate its power and efficiency. The command line interface, while initially daunting, allowed me to perform complex tasks and automate repetitive processes

quickly and easily. I found that by learning how to use Linux effectively, I was able to significantly streamline my workflow and increase my productivity.

The use of AXAEM, an archival management software, also presented some significant challenges during the development of the Pre-Ingest Processing Manual for Digital Content. The software, while powerful and feature-rich, was not always easy to work with and often required a significant amount of problem-solving and consultation with the vendor to overcome issues and errors that arose.

One challenge that I encountered when working with AXAEM was that error reports often did not provide sufficient information to indicate the frequency of the error. This meant that I had to rely on the archives staff to identify and track these errors manually, which was time-consuming and sometimes difficult. This challenge required us to be proactive in our approach, and we often had to pivot and change our methods to move forward and overcome the problem.

Another challenge we faced when working with AXAEM was the vendor's response to issues that we identified. In some cases, the vendor did not provide a sufficient remedy to the problem, which meant that we had to find workarounds or alternative solutions to address the issue. This required a significant amount of problem-solving and collaboration among the archives staff, as well as consultation with the vendor to ensure that we were using the software correctly and effectively.

Despite these challenges, I was ultimately able to develop a Pre-Ingest Processing Manual for Digital Content that was comprehensive and effective. My experience working with Linux and AXAEM helped me to better understand the complexities of processing and ingesting digital materials into an archive's digital repository, and I was

able to develop new strategies and techniques that allowed me to overcome these challenges and move forward with the project.

CONCLUSION:

This thesis paper has explored the challenges posed by digital preservation and the importance of developing a pre-ingest processing manual for digital content. The Open Archival Information System (OAIS) reference model has been widely accepted and adopted in the digital preservation field, but it provides little guidance in the practical implementation of digital archival preservation workflows. This has resulted in the development of various methods and tools to comply with the OAIS model, but there is inadequate documentation of these methods in the scholarly literature.

The project described in this paper at Duke University Medical Center Archives offers a practical solution to this problem by enhancing the repository's current Electronic Records Processing Guide. The revised guide was developed and tested using the digital materials from two recent digital accessions to the Garnett H. Kelsoe Laboratory Notebooks and the Physician Assistant Program Records. The outcomes of the pre-ingest processing manual for digital content will provide increased stability in the transfer of electronic records from digital carriers and efficiency in processing a larger volume of digital materials by utilizing open-source digital forensic tools to process them for ingest into the repository's OAIS-compliant integrated archives management system.

This project has contributed to the ongoing effort of preserving born-digital materials by providing a tangible solution to the challenges of digital preservation. The pre-ingest processing manual will serve as a valuable resource for archivists and

institutions seeking to implement OAIS-compliant digital archival preservation workflows. Moreover, this project underscores the need for continued research and documentation of practical methods and tools for digital preservation, which will facilitate the long-term preservation and accessibility of digital materials for future generations.

WORKS CITED:

- Arroyo-Ramirez, Elvia, et al. “‘Tell Us About Your Digital Archives Workstation’: A Survey and Case Study.” *Journal of Contemporary Archival Studies*, vol. 5, 2018, <https://elischolar.library.yale.edu/jcas/vol5/iss1/16>.
- Barrera-Gomez, Julianna, and Ricky Erway. *Walk This Way: Detailed Steps for Transferring Born-Digital Content from Media You Can Read In-House*. OCLC Research, 2013, <http://www.oclc.org/content/dam/research/publications/library/2013/2013-02.pdf>.
- Baucom, Erin. “A Brief History of Digital Preservation.” *Mansfield Library Faculty Publications*, 2019, https://scholarworks.umt.edu/ml_pubs/31.
- Belovari, Susanne. “Expedited Digital Appraisal for Regular Archivists: An MPLP-Type Appraisal Workflow for Hybrid Collections.” *Journal of Archival Organization*, vol. 16, no. 4, Oct. 2019, pp. 197–219. *Taylor and Francis+NEJM*, <https://doi.org/10.1080/15332748.2019.1682793>.
- . “Expedited Digital Appraisal for Regular Archivists: An MPLP-Type Approach.” *Journal of Archival Organization*, vol. 14, no. 1–2, Apr. 2017, pp. 55–77. *Taylor and Francis+NEJM*, <https://doi.org/10.1080/15332748.2018.1503014>.
- Chassanoff, Alexandra, and Colin Post. *OSSArcFlow Guide to Documenting Born-Digital Archival Workflows*. Educopia Institute, 2020, https://educopia.org/wp-content/uploads/2020/06/OSSArcFlow_Guide_FINAL-1.pdf.

- Cirella, David. *Understanding Bulk Extractor Scanners*. Confluence, 2020,
<https://confluence.educopia.org/display/BC/Understanding+Bulk+Extractor+Scanners>.
- Consultative Committee for Space Data Systems. *Reference Model For An Open Archival Information System (OAIS)*. June 2012,
https://en.wikipedia.org/wiki/Consultative_Committee_for_Space_Data_Systems.
- Corrado, Edward. "Digital Preservation Is Not Just A Technology Problem." *Technical Services Quarterly*, vol. 39, no. 2, 2022, pp. 143–51,
<https://doi.org/10.1080/07317131.2022.2045432>.
- DeRidder, Jody L., and Alissa Matheny Helms. "Intake of Digital Content: Survey Results From the Field." *D-Lib Magazine*, vol. 22, no. 11/12, Dec. 2016,
<https://doi.org/10.1045/november2016-deridder>.
- Durno, John. "Digital Archaeology and/or Forensics: Working with Floppy Disks from the 1980s." *Code4Lib Journal*, no. 34, Oct. 2016,
<https://journal.code4lib.org/articles/11986>.
- Erway, Ricky. *Defining "Born Digital"*. OCLC Research, Nov. 2010,
<http://www.oclc.org/research/activities/hiddencollections/borndigital.pdf>.
- Eschenfelder, Kristin, et al. "What Are We Talking about When We Talk about Sustainability of Digital Archives, Repositories and Libraries?" *Proceedings of the Association for Information Science and Technology*, vol. 53, no. 1, 2016, pp. 1–6.

File Profiling Tool (DROID). The National Archives,
<https://www.nationalarchives.gov.uk/information-management/manage-information/policy-process/digital-continuity/file-profiling-tool-droid/>. Accessed 16 Apr. 2023.

Goldman, Ben. "Bridging the Gap: Taking Practical Steps Toward Managing Born-Digital Collections in Manuscript Repositories." *BM: A Journal of Rare Books, Manuscripts, and Cultural Heritag*, vol. 12, no. 1, Mar. 2011, pp. 11–24,
<https://doi.org/10.5860/rbm.12.1.343>.

Johnston, Leslie. "Challenges in Preservation and Archiving Digital Materials." *Information Services & Use*, vol. 40, 2020, pp. 193–99,
<https://doi.org/10.3233/ISU-200090>.

Niu, Jinfang. "Appraisal and Selection for Digital Curation." *International Journal of Digital Curation*, vol. 9, no. 2, 2014, pp. 65–82.

Sloyan, Victoria. "Born-Digital Archives at the Wellcome Library: Appraisal and Sensitivity Review of Two Hard Drives." *Archives and Records*, vol. 37, no. 1, Apr. 2016, pp. 20–36, <https://doi.org/10.1080/23257962.2016.1144504>.

Walsh, Tim. "Data-Driven Reporting and Processing of Digital Archives with Brunnhilde." *Practical Technology for Archives*, no. 8, 2017,
<https://hdl.handle.net/1813/76867>.

Woods, Kam, and Christopher A. Lee. *Acquisition and Processing of Disk Images to Further Archival Goals*. 2012, <https://ils.unc.edu/callee/archiving-2012-woods-lee.pdf>.