# Forensically Enhanced Digital Preservation

by

**Timothy Robert Hart**

*Thesis*

*Submitted to Flinders University*

*for the degree of*

**Doctor of Philosophy**

College of Science and Engineering

16/05/2022

"I certify that this work does not incorporate without acknowledgment any material previously submitted for a degree or diploma in any university; and that to the best of my knowledge and belief it does not contain any material previously published or written by another person except where due reference is made in the text."

*Sign* Timothy Robert Hart *Date* 16/05/22

# TABLE OF CONTENTS

# LIST OF FIGURES

## LIST OF TABLES

## Dedication

This thesis is dedicated to all whom have supported me throughout my years of study. To my partner, my family, my supervisor, and all the mentors and teachers that saw something in me, you all made this possible.

To anyone who has doubted themselves or have been doubted by others, perseverance is the key, and in the end it is worth it.

## Acknowledgement

First and foremost, I acknowledge Denise de Vries who has been a wonderful supervisor, mentor, and friend. For with her support and guidance I was able to learn so much through this candidature. Our fortnightly meetings always helped keep me confident and on track. Denise always took the time to review drafts and provide extensive feedback which greatly benefitted this thesis. I could not have asked for a better supervisor.

In no particular order, I must acknowledge my partner, my parents, and my immediate family who always have supported and believed in me.

My partner supported me and put up with me through this stressful time. She always gave me confidence and made me believe in myself and my work.

Without my parents I would not have had the opportunities that led me to this point. I was able to study without any burdens, allowing me to dedicate all my time towards achieving my goals.

The pride from my family has always been a great motivator and I will forever be thankful for their love and support.

To all my mentors, teachers, and lecturers that helped guide me towards academia, I thank and acknowledge you.

A special thanks to Anna Shillabeer for her help and support.

This research was made possible by the Australian government research training program scholarship, the Australian Postgraduate Award – 2016.

## Publications

Metadata Provenance and Vulnerability, Hart, T.R., D. de Vries, Information Technology and Libraries, Vol 36, Issue 4, P 24-33, 2017, DOI 10.6017/ital.v36i4.10146

Australian Law Implications on Digital Preservation, Hart, T.R., D. de Vries and C. Mooney, iPres 2019, Amsterdam, Sep. 16, 2019, DOI 10.17605/OSF.IO/EZ6FQ

## Abstract

Digital preservation and digital forensics are two fields with differing goals that travel similar pathways which often converge. Each field does not necessarily acknowledge the other, but they are closely aligned and share similarities. Digital preservation has much to benefit from digital forensics; however, this is not to say digital forensics could not gain with respect to documentation and perspective with collaboration in mind.

One of the key differences is long-term preservation, where the material is stored and maintained long after it has been processed versus forensic evidence gathered and used to prosecute, with no further regard once done so. The efforts that go into ensuring the preservation of digital objects are where the similarities between the two fields end. This results in digital forensic software being tailored to the specifics of the field, such as modern devices, specific data, and criminal prosecution. Perspective and purpose are important factors as they determine how the software is perceived and documented. This affects the adaptability of digital forensic software for memory institutions (galleries, libraries, archives, museums) as at face value, it does not cater to their needs, despite being beneficial.

In this thesis, the benefits of using digital forensic software for born-digital preservation are explored, as well as the risk to collections should data remain unprocessed via the suggested methods.

Hidden data may already exist within storage collections, yet to be discovered and impossible to do so without the use of digital forensic software. These data, rightly named "sensitive data" have many implications. Sensitive data, whilst the key to criminal investigations, is also paramount to digital preservation as it can reveal significant amounts of new information.

Australian law is explored regarding the risk of sensitive data discovery and the actions that follow. The threats of sensitive data are discussed with consideration to the potential legal implications that arise from discovering sensitive data. This includes examples of current and future threats that may reside in stored data that have not been processed assiduously using digital forensic software.

Policies and procedures regarding Aboriginal and Torres Strait Islander people and their information are explored and compared against standard policies. The strict and careful policies developed for our Indigenous people can positively influence the standard privacy policies within institutions implementing or advancing sensitive data discovery.

The scope of this study has been narrowed down to Australian institutions, targeting State and National libraries whilst also considering archives, galleries, and museums, as these are the influential institutions. Australian institutions have been investigated by the information publicly available and by communications, distributing a questionnaire to willing participants.

Collection institutions from the United States of America were investigated to form a comparison and to establish potential tools and methods that could be adopted within Australian institutions. The data gathered from the U.S institutions were derived by publicly available information and other studies conducted. The main sources of data were derived from workflows as these allowed a visual representation of the processes and the tools used within collection institutions, revealing if and where digital forensics was being utilised.

It was evident the major collection institutions of Australia were performing digital preservation at various maturity levels. Intake requirements and dedicated preservation procedures were varied, as was the influence of digital forensic tools and methods.

Some of the participants of the study identified the need for improvement regarding their workflows, whereas others had low demand and therefore did not see the need to make any changes. It was determined that some digital forensics was being utilised, but not to its full potential, and in most cases, was missing completely. With the analysis of collection institutions and the benefits of digital forensics, the objective is to increase awareness and provide workflow improvements to enable sensitive data discovery and the handling of any surrounding issues that may arise.

The identification of maturity levels for digital preservation in Australian institutions has been established by the feedback provided via questionnaire and data gathered from public sources. This information, compared with other institutions and maturity level modelling allowed the establishment of an average baseline in terms of maturity levels of digital preservation requirements and performance.

Digital forensic tools and methods have been analysed to determine the data gathering capabilities of digital forensic software and the relevance to digital preservation. The benefits of digital forensics within digital preservation workflows and the impact of sensitive data within collection institutions form the contributions of this study.

Experiments have been conducted in real world scenarios using donated material (hard drives), resulting in a plethora of data gathered with an extensive range in severity. The

potential for sensitive data discovery was revealed as well as the ability to derive information about the users of the physical media.

Issues regarding digital preservation workflows have been identified.  Many workflows are missing core processes that are required to handle sensitive data. This may be the result of either a lack of transparency, where sensitive data discovery is being performed to some extent but is undocumented, or the process is missing entirely.

Through the process of reviewing and analysing workflows, good practices were also identified, resulting in the discovery of exemplary workflow designs to help in determining how digital preservation workflows can be improved.

Amendments and enhancements to workflows to address sensitive data discovery are presented, enhancing digital preservation workflows with digital forensic tools and methods. This is not only to improve existing institutions, but also to better enable peer-to-peer learning and collaboration.

With the implementation of digital forensics within mature and influential collection institutions, other institutions that may be in their infancy or slowly developing their procedures will have guidance. This can be achieved with transparent workflows that accurately visualise the forensic processes, addressing all outcomes and decision-making, and documenting the tools used and any implementation requirements.

# 1 INTRODUCTION

Digital preservation as a field of endeavour is gaining traction and is slowly being recognised for its importance and necessity. However, this may not always be the case in communities outside of the discipline. Whilst preservation is open to adopting other disciplinary methods and techniques, other related disciplines may not be considering preservation when designing tools or publishing literature. If preservation were a consideration within these fields, it could potentially influence how they develop and publish, which in turn would make the benefits they offer more discoverable and easier to adopt for collection institutions. Collection institutions are also commonly known as memory institutions, although this term encompasses a broader range of organisations which maintain a repository of public knowledge. Whilst it would be great to have the importance of digital preservation recognised within other related fields, the best approach is to focus on improving digital preservation, thus increasing its chance of exposure.

The institutions in which digital preservation is typically performed are not always equipped to handle the tasks adequately. This includes available resources and trained personnel. One cannot assume librarian staff, for example, have the technical background and knowhow to be aware of solutions that could aid them in their work that fall outside of their discipline. The overarching issue of awareness is something being worked on as a community through conferences and other group activities. Gatherings of like-minded digital preservation enthusiasts band together, sharing their discoveries, challenges, and solutions with the proceedings published for all to read.

However, being aware of the outcomes and deliverables that come from these gatherings may not always reach the smaller institutions or countries that do not participate in these events. This is one of the reasons why some collection institutions are performing at different levels of maturity as they may not be aware of better solutions.

Regarding different levels of maturity, the reasons behind this may differ. Some institutions are quite far along in their development, whilst others may be in their infancy. Some institutions may have procedures and policies in place for every stage of a digital preservation lifecycle, including adequately designed workflows. Some may only have procedures in place to handle intake, with no dedicated preservation methods to handle the remaining tasks. There may be some institutions with little intake and therefore no dedicated preservation methods are required. This suggests the maturity level can be tied to the amount of material requiring preservation within these institutions. Evidence of this is provided throughout this study.

One thing is certain, this is an evolving field, and this means there are improvements to be made in all areas. With institutions being at different levels, collaborative and cascading learning results from there often being partnerships between institutions and groups of institutions that form a larger collaborative organisation. Whilst this forms part of the solution, that being, if improvements are made in one institution, the peers have a good example to follow and learn from; this also reveals a major issue if incomplete, inaccurate, or obfuscated examples are followed.

Transparency is a term used frequently in this study and it is the concept of collection institutions being completely transparent in their digital preservation workflows. This means every process should be visualised to show how data is handled during and after these events occur, as well as any decision-making conducted to direct these data and processes. Without this transparency, smaller institutions do not have a readily available, and appropriate, guide to follow when adopting new practices. It then falls on their awareness of existing solutions and their ability to research potential solutions, in which they are likely to look for guidance from their peer and partnered institutions.

Lack of transparency may occur for several reasons. An institution may not have dedicated procedures to visualise, and the preservation being performed may be done in an ad hoc manner. An institution may also wish to hide their preservation techniques from other institutions should there be competitive factors involved. The staff within the institution may not have the required knowledge to accurately visualise and develop workflows. Whatever the reason is for the omission of certain processes, whether they are being performed and not visualised, or not being performed at all, a proper transparent workflow can reveal this information, leading to improvements and further development.

Whilst these issues are focused on the overall improvement of digital preservation across multiple institutions, there are risks and issues for individual institutions that may already be present and are certain to be sustained in the near future. This is where the issue of sensitive data is introduced. The adequate discovery of such data may not be possible in many institutions that have not adopted any digital forensic tools and methods in their digital preservation processes. Furthermore, it cannot be assumed that sensitive data are handled appropriately in intuitions that are partaking in digital forensics. Sensitive data bring risk to collection institutions, legally and ethically, but may also strengthen collections with accuracy and completeness of collection items.

Digital forensics adheres to forensic sciences applied to law where the principles, methodologies, and techniques are used to aid in forensic investigations. (Sachowski, 2016, p. 1). Digital forensics has evolved overtime with evolution of cybercrime. The tools and methods available through this discipline enable capabilities within collection institutions to better handle digital devices and born-digital data.

> *"From the traditional computer system to modern devices such as mobile phones, game consoles, or virtualized environments, the field of digital forensics encompasses a wide range of technologies that serve as potential evidence sources. While the design and functionality of these technologies is uniquely different, the application of digital forensics involves ensuring the integrity and authenticity are upheld throughout the evidence's life cycle."* (Sachowski, 2016, p. 1)

There lies the risk in unprotected data that are sitting in storage or have been discarded. These data have not been completely evaluated for potential information, both useful and harmful to the institution. The way data are ingested into a collection plays a big role in this. Donations of digital artefacts present the greatest threat of sensitive data, as do the curation of computing systems and hard drives. These media contain extensive amounts of data, hidden in obscure locations, only retrievable by the appropriate digital forensic tools and methods.

> *"With very large archives, sensitivity review can be a Sisyphean task. It is always possible that collections that were deemed non-sensitive turn out to contain problematic materials."* (Jaillant, 2022)

Therefore, as intake plays a large role, this may be unprecedented in small institutions. The type and quantities of data ingested will determine the nature of sensitive data discovered. This is something that will grow, undoubtedly. As this growth occurs, the risk increases. With new issues, come new solutions, and these solutions often present new issues. Legal issues, ethical decision-making, and resource limitations are the major concerns raised when adopting solutions to handle sensitive data.

As sensitive data forms the basis and defines the issues this study aims to resolve, within Australian collection institutions, the ways in which digital forensics can aid in fixing and improving them are explored. The overall issue being faced is the discovery and handling of sensitive data. Whilst the focus is on Australian institutions, many of the discoveries and solutions provided in this study are relevant to all collection institutions with preservation goals, despite their maturity levels and use of digital forensics.

The first investigation into sensitive data is based on how “sensitive data” is defined, the effect it can have, and what issues surround it. With sensitive data defined and understood, the investigation looks towards Australian law and how it relates to such data. When dealing with collection institutions, such as libraries, archives, galleries, and museums, which may also reside in universities, complexities arise as exemptions from privacy law protect the collections. It is after all the task of the collection to provide accurate and honest information about our history. If this were to be enforced and collection institutions were obligated to follow standard privacy law, it would severely hinder the capabilities of the collection.

Exemptions, however, do not eliminate the presence of legality issues. Therefore, privacy law has been considered and investigated. Due to this, ethical considerations are also investigated as the majority of the decisions made around sensitive data will be discussed and made, based on ethical and moral standards.

Australia’s Indigenous presence in collection institutions is explored, including the unique legal and ethical issues for Indigenous material. Whilst the specifics may be unique to Australia’s Aboriginal and Torres Strait Islander people, there may be similarities in practices for other indigenous cultures, therefore, making it a worthwhile consideration outside of Australia.

With the importance of sensitive data established, the focus is directed towards the discovery and handling of these data. Given the sensitive nature of digital forensic investigations, in which sensitive data and metadata are essential discoveries, it is determined that the same approach can be taken to suit the needs within collection institutions. Digital forensic tools and methods allow data to be discovered in obscure locations, unreachable by manual means, and within a fraction of the time it would take to manually conduct this process. This reduces the resources required in search and retrieval tasks, allowing efficient allocation of resources to the analysis of the output provided by digital forensic tools, typically presented in a user-friendly manner that is easier to interpret.

Before the exploration of digital forensic tools and methods, investigations were conducted on international and Australian collection institutions. This involved research into online and public documentation made up of websites and published literature. The focus was on workflows and the data that could be derived from them such as what tools were used within the digital preservation process. Australian institutions were presented with a questionnaire aimed to establish all the required and relevant information to determine their working procedures and assess their maturity level of preservation. The presence of digital forensics implementation and consideration was investigated when analysing the public information.

Therefore, the four primary areas of focus are: The maturity of digital preservation in collection institutions, the types of preservation tools used, the use of digital forensic tools specifically and digital preservation workflows within those institutions.

This presented data that could be calculated and compared, resulting in discoveries of which tools could be potential candidates for broader experimentation. It also raised questions on why certain tools were used, within varying circumstances. Comparative data allowed views on influential tools, used across multiple institutions, and revealed many unique tools that only had one occurrence across the datasets. The correlation of such data with the accompanying information provided via the questionnaire allowed patterns and inconsistencies to be revealed, all of which helped establish a better understanding of Australia’s progress in digital preservation and the tools and methods used to accomplish preservation goals.

Digital forensic tools were then specifically investigated and utilised against real data in a real environment. The data came from donated media in which the content was unknown to provide an authentic experience. The appropriate steps were taken to image and assess the media against criteria set to determine its testability. The data from the digital forensic tool output were analysed and presented from a digital preservation perspective, focusing on data that could both hurt and help a collection institution regarding ongoing meaningful access to preserved data. The features of the tools that could aid investigations were explored and documented. The use of the output data, correlated with multiple findings, displayed how these data can be used to profile an individual, something that may be beneficial when dealing with iconic figures of history and persons of interest.

Sensitive data can reveal and expose people or groups, in both a positive and negative manner. The primary goal of a collection is to provide authentic, accurate, and complete information to the public. However, the potential of digital forensic software may alter this point of view as there are cases where the omission of information is necessary. Some information may be in the best interest of the public and should therefore be protected against legal and ethical concerns, but there are instances where protection may be questionable.

Sometimes information should be kept from the public if it serves no purpose other than to hurt the reputation of others. This form of decision-making has been considered when developing solutions in the form of workflows, the next area of focus.

Workflows provide an overview of the preservation processes in varying levels of depth. Digital preservation treats workflows differently from other disciplines as they are typically used as guidelines, often flexible and changing. The nature of digital preservation and how each case is unique is why workflows are not definitive procedures for a system to follow, especially given the user input requirement. The human element plays a large role in this discipline.

Issues were discovered surrounding digital preservation workflows. Based on reviewed documentation and existing workflows, transparency was not achieved at a high enough level as many critical processes were not being visualised. Therefore, the goal was to take what was learned from digital forensics, apply it to digital preservation, and visualise it within workflows for collection institutions to follow. By being transparent with this approach, other peer institutions that may not be on the same level of maturity will have an exemplary model to follow when their preservation needs are increased. This overall leads to better digital preservation practices across the country.

The workflows presented in this study improve the already existing initial stages of digital preservation, such as the donor agreement phase. The main process of preservation is improved by the core enhancements developed that include the implementation of digital forensic tools and methods, along with extensive decision-making to ensure no data can pass through the workflow that may pose a risk without a proper evaluation. Sub-diagrams have been provided for certain processes to reduce complexity and make the workflows more manageable. The solutions provided have been presented in a way that is free of notation and design, allowing institutions the choice in how they adopt the enchantments. Design ideas have been reviewed and presented. Options have been considered throughout the design process, with resource limitations in mind.

At the conclusion of this study, a discussion is presented that culminates each chapter and describes the overall process. In this discussion, the options that will be presented before the institutions that choose to adopt these enhancements are explored. The considerations made when developing the solutions are discussed as are the requirements for implementation.

The remainder of this chapter establishes the research questions, the aims and objectives, and defines the scope of this study.

## 1.1 Research Questions

The questions asked and the drive behind the following research is made up of four subject areas: digital preservation, digital forensics, ethics, and legal considerations. The issues and solutions discussed flow from one another. The digital preservation issues explored can be resolved with digital forensic tools and methods, however, with the implementation of digital forensics, legal and ethical issues arise. When there are no legal solutions, ethical and moral based decision-making must be conducted as collection institutions are in a unique position where they may legally publish information but does not mean they should. Each research question presented is considered high-level, encompassing other questions that may stem from them, and which are typically discovered over time through research and experimentation. The questions cover all areas necessary to achieve the goals of this research of which are to enhance digital preservation with digital forensic tools and methods, enabling the discovery and handling of sensitive data. These enhancements will help prevent legal and ethical issues whilst strengthening collections with new, previously undiscoverable data, that adds to the accuracy and completeness of collection items.

The first question is based on digital preservation, the main subject that encompasses the other subjects:

> **Question One** – *Where can improvements and amendments be made, and are they required, in current and future digital preservation workflows to allow for greater data gathering capabilities in Australian collection and memory institutions?*

This question addresses workflows and considers the working procedures of the institutions as well as the workflow diagrams and visualisations. These are two crucial elements as one reflects the other, with the diagrams being a public reflection of the institution's digital preservation strategy, which is important for transparency, a subject stressed throughout this study. The improvements identified and presented here address the use of appropriate digital forensic tools and techniques to facilitate in the discovery of sensitive data and other useful metadata that may aid collections with preservation and descriptive based context.

> **Question Two** – *How can digital forensic tools and techniques be implemented to resolve the data gathering issues existing within collection institutions where data hidden in obscure locations is not being addressed?*

By addressing this issue, collection institutions will be able to adopt new methods and techniques that allow them to access a plethora of new data as well as gaining greater value from unrealised information hidden within current data. The information discovered via these methods can aid collections in a better understanding of the data which they hold. It may provide new information on a subject, system data to aid in emulation, and provenance data that can help in establishing the digital history of collection items.

With these capabilities, new issues potentially emerge. Introducing new methods, tools, or techniques, both invasive and thorough in nature, will inevitably lead to legal and ethical complications.

The following two questions address if there are procedures and policies in place to handle the issues that arise from the use of digital forensics and the discovery of sensitive data. These are categorised as "legal" and "ethics".

> **Question Three** - *Are legal implications considered that would be in effect if not for State and National institution exemptions with respect to Australia's Privacy Principles?*
>
> **Question Four** - *What are the ethical procedures in place and how is the decision-making process conducted when dealing with sensitive data that have no legal concerns, but may be considered in an ethical or moral grey area?*

Whilst collection institutions have the primary goal of providing access to accurate and complete information, outside of copyright and embargoed restrictions, privacy is typically not a concern. These institutions are not subjected to the same privacy laws as other institutions. Questions three and four address how institutions treat privacy policies, despite their exemptions, together with how decisions are made when ethical and moral issues are of concern. The issues from which these questions are derived from are explored in Chapter 4 AUSTRALIAN LAW IMPLICATIONS.

## 1.2 Aim and Objectives

The main sources of data for this study were extracted and analysed from workflows and public online documentation. Workflows are often visualised as activity-diagrams that show the flow of processes and the users within a system, in this case, digital preservation from ingest to storage. Each node within these workflows represents an action, performed by users, systems (software and hardware), or both.

The benefit of analysing workflows, or conceptually constructing them based on gathered information where an existing workflow may not be available is useful. With workflows, the tools being used within an institution for their digital preservation process are visualised, another key bit of information used in analysis. It also reveals missing processes as the diagram can be followed systematically to detect any vulnerabilities. This can range from a process missing entirely to a lack of error handling.

If sensitive data are discovered, there should be a process to handle it with considerations to legal and ethical standards, and it should be visualised in the workflow. However, this may not always be the case and there are instances where a workflow may not visualise post-discovery handling. Handling involves any decision-making regarding the data, the use of it, and the procedures in place for the data deemed unfit for ingest. The following is an example of a basic decision process in a workflow:

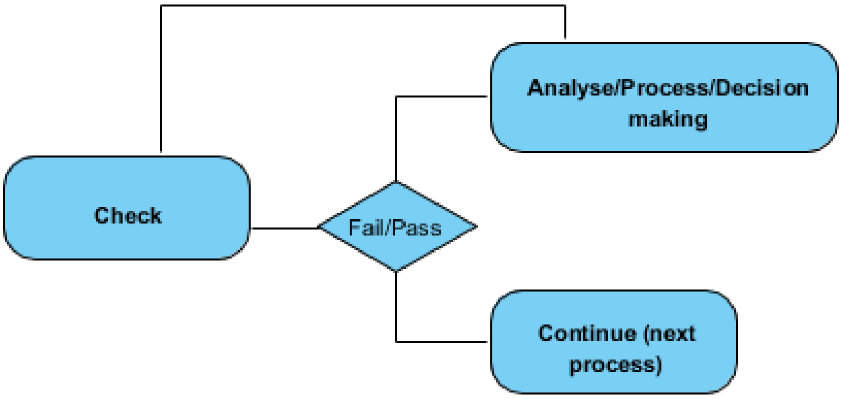


**Figure 1- Decision Diagram**

The objective is not to completely change existing workflows, but to focus on areas believed to need improvement and additions. Figure 1 is an example of an improvement that could be made for certain processes. Error handling such as this is often utilised, but it is sometimes overlooked. There are cases where decision nodes should be present, ensuring the process cannot proceed unless certain criteria are met. Without such safeguards, the workflow proceeds to the next node, regardless of the resultant state of the previous process. The workflows are intended to be used as guides which are flexible and can be adapted to each unique case rather than a prescribed set of steps that must be followed systematically.

It is not suggested that error-handling is lacking when such processes are not visualised. The pertinent point is that these steps are not included in workflows. To avoid any "point of failure" a workflow should be complete and accurate. This is important for training purposes as well as knowledge sharing. Accurate, complete, and transparent workflows can provide invaluable assistance to those who are in the early stages of digital preservation.

To summarise, workflow improvements via the implementation of digital forensics to enable sensitive data discovery is the main objective with the additional aim to improve ethical decision-making whilst maintaining legal consideration when not obliged to. This leads to reduced risk, better and accurate collections, and overall improvements. This can then result in collaborative learning within partnered collection institutions, such as those within the National and State Libraries Australia (NSLA), and should transparency be achieved, institutions outside of partnered circles can learn from their example, eventually progressing digital preservation throughout Australia.

Therefore, the key objectives and aims are:

- Enabling sensitive data discovery with digital forensic tools and methods
    - Raising awareness of the potential of sensitive data (risks and benefits)
- Ensuring sensitive data are handled appropriately once discovered (with ethical and legal considerations)
- Accurately reflecting these objectives in workflows
    - Achieving transparency in workflow processes and tools to promote collaborative learning

## 1.3 Scope

The terminology used within this study regarding digital preservation is used with a broad perspective on the subject. When referring to "digital preservation", the life cycle of digital preservation is being assumed and not just the act of preserving digital objects. This includes pre-acquisition through to storage and access.

Pre-acquisition includes the procedures in place for accepting donations and how donor-agreements are formed. Once donations are accepted, the processes to identify and handle sensitive data are of concern. Sensitive data discovery may be a new addition, or improved from an existing process, making use of digital forensic software to extract sensitive data from donated material. The use of or lack of digital forensic tools and methods give insight into whether sensitive discovery is performed, or the level at which it is performed.

Legal issues that arise from the discovery of sensitive data are considered within the scope of Australian privacy law. Other jurisdictional laws are not considered, however, examples of ethical issues outside of Australia are deemed relevant.

The act of preserving digital material is not within the scope of this study. No changes are being suggested to the core preservation workflow to ensure no major disruptions occur should any of the suggested enhancements be adopted.

The remaining stages of concern are storage and access. This includes temporary storage, acting as a buffer between stages of processing or where material may sit idle whilst ethical and legal decision-making occur, and final storage, where material is maintained and access is provided.

The National Library of Australia defines the primary objective of digital preservation activities as:

> *"maintaining the ability to meaningfully access digital collection content over time. The primary concern is preserving the ability to access the Preservation Master File from which derivatives files may be created or re-created over time. To this end, preservation of digital library material includes:*
>
> - *Bit-level preservation of all digital objects which means keeping the original files intact;*
> - *Ensuring that authenticity and provenance is maintained;*
> - *Ensuring that appropriate preservation information is maintained;*
> - *Understanding and reporting on risks which affect ongoing access;*
> - *Performing appropriate actions on sets of digital objects to ensure that the objects continue to be accessible; and*
> - *Periodic review of preferred formats and digital metadata standards"* (NSLA, 2013)

Therefore, the stance taken is to ensure preservation is improved by ensuring the acquisition stages are done so with consideration to the suggested enhancements and the information provided on sensitive data. This will ensure meaningful access can be achieved.

The scope of this study was narrowed to only target Australian institutions. This was decided for two important reasons. The first reason is the varying legislation and ethical ideologies that differ across jurisdictions, making a worldwide approach less achievable.

The second reason for narrowing the scope to Australia is the necessity for improvements in digital preservation, an assumption that was confirmed via communication with the institutions in question.

Initially, the scope included the larger state and national institutions of galleries, libraries, archives, and museums (GLAMs) as these are of great importance and influence for their residing states. Upon communications with various institutions, it became clear that libraries were more approachable, therefore, the scope was narrowed down to national and state libraries of Australia. There were limitations in the final sample size as some institutions were unable to participate, stating they were not developed enough to provide adequate information, and others did not agree to participate or had to withdraw for unspecified reasons.

The current objectives within libraries better align with this study, however, this does not exclude the other institutions benefiting from this research. This decision was made based on the feedback received in response to the questionnaire as well as time spent at one of the Australian state galleries where several institutions within Australia attended, made up of libraries, archives, and galleries. The attendees from libraries showed greater interest in the sensitive data aspects of this research, whereas the museum and archives showed greater interest in the ability to capture more meaningful data regarding their subjects. Libraries, however, were more flexible and open to changes, which is a factor in why they were chosen as the primary target.

Among libraries, national and state libraries have more exposure, a public presence, and are considered influential. Therefore, if improvements are adopted within these institutions, it will influence others. With this notion, it was concerning upon initial investigation that the publicly available information and transparency regarding the digital preservation process and related polices were lacking. The gaps identified revealed the need for improvements, improvements which would be widely influential given the status of these institutions.

### 1.3.1 Limitations

Limitations on the output of this study were considered for various reasons. According to the correspondence with participants of this research, it is clear these institutions are not able to adopt drastic change that will impact current working order. Therefore, the scope of the solutions suggested has been carefully focused on enhancements that can significantly improve existing workflows without changing any core processes and allowing implementation at the user’s discretion. Institutional resource limitations are considered; therefore, any solutions suggested have been kept within the scope of needing to be readily accessible and open-source, allowing freedom of choice.

Datasets were derived from institutions based in the U.S as they met the requirements of being publicly available (transparency) with an influence of digital forensics. The first two sets from the BitCurator consortium provided a comparison of workflows from 2012 to 2016, all which were influenced by BitCurator and digital forensics. The third set from Educopia Institute’s OSSArcFlow (Open Source Software Archival Workflow) provided a more recent set of workflows within similar characteristics. The limitations of the data available publicly and to be provided by Australian institutions resulted in the need to look to exemplary models to form comparisons and influence the new workflows presented within this study.

The targeted institutions in this research were the national and state libraries and archives which resulted in a small sample size. However, these institutions represent the entire area of Australasia. The states and territories include South Australia, Tasmania, New South Wales, Victoria, Western Australia, Northern Territory, Queensland, Australian Capital Territory, and New Zealand.

Limitations arose in the participation amongst these states. Some institutions could not participate based on their current maturity level; a concept explained in Section 2.6 Maturity Levels. There were also withdrawals from participation as well as issues maintaining correspondence. This was prominent in the state and national archives approached, of which the returned data was not satisfactory enough to be used in any analysis.

It should be noted that the lack of data in certain areas is in itself data. The identification of gaps and areas that need improving helped narrow the focus of this research.

Further limitations were imposed to meet the end goal of this research in a manner more likely to be adopted by the institutions to which this research relates. This includes the consideration of resources, specifically budgetary and staff restrictions. Therefore, cost and complexity of solutions impacted the tools and methods investigated and proposed.

## 1.4 Motivation

Primarily, the motivation behind this study is the betterment to all stages of digital preservation. Equally as important is the increase of awareness and recognition amongst other disciplines regarding how they can add benefit to digital preservation. With this recognition, disciplines such as digital forensics can start to consider additional applications of their tools and methods. This may then lead to consideration towards other fields such as when developing the documentation for tools developed in such disciplines. Perhaps support from developers and providers may then be provided for these alternative applications.

Anyone dealing with or creating data that will eventually make their way to a collection institution for preservation can also help in this regard. If they recognise the importance of keeping their data alive and secure, they may take steps to ensure the metadata are kept accurately, such as any provenance and change history, which in turn will make it easier for the collection institution to preserve their data. They may also consider how their data are stored and the media on which they are stored. Overall, any extra care and precautions during the creation and acquisition of born-digital data will increase the effectiveness of digital preservation.

Recognition and awareness will provide the means to advance digital preservation practices towards solving existing problems and preparing, therefore, preventing, future issues.

It is a strong personal belief that being prepared for the future and preventing issues are critical practices. Solving problems as they arise may not allow all damage to be mitigated or repaired. Therefore, with the investigations and solutions presented within this study, a better understanding of these issues can be achieved. With this understanding, more appropriate preservation preparations can then be made.

The criminology side of digital forensics, and the nature of digital investigations based on criminal activity has been a strong motivator. The extent of data gathering capabilities and the ability to discover and recover data from obscure locations are ideal for what needs to be achieved. The ability to identify and correlate patterns of information to discover new and interesting facts about a subject of interest strongly benefits collections.

With these principles, collection institutions can:

- prevent legal and ethical issues,
- discover environmental data (system, devices, software, hardware),
- discover contextual information (user information, hobbies, interests, political views, etc.),
- and significantly reduce time needed to analyse, sort, and categories data.

The information derived from data and the knowledge gained by making this information available to the public will be improved by having more data that are accurate, complete, and trustworthy.

It is in the public interest that history, noteworthy events, and related subjects are known. It is desirable that, as near as possible, complete information is made available. Progress towards this goal can be achievable with the influence of digital forensics, enabling better handling of data and their media.

As each issue is addressed and more solutions are provided, the institutions that take on this responsibility to improve are then driving digital preservation one step closer to fixing and preventing issues on a global scale. If more institutions adopt digital forensic methods and apply them appropriately, there will be more exposure for the necessity of such implementations. This may eventually lead to the two disciplines addressing their similarities and aligned goals, bringing them closer together which can result in better collaboration, tools, methods, and each field being more considerate of one another.

## 1.5 Organisation

The organisation of this thesis is as follows.

Chapter 2 LITERATURE REVIEW provides an overview of related research of digital preservation, digital forensics, and how their goals compare with respect to born-digital data, including the media on which they are stored. The growth of data and the increase in digital preservation requirements are investigated.

The areas of focus are:

- A comparison of digital preservation and digital forensics
    - Ethics, privacy, and legal comparisons
- Provenance
- Digital forensic methods and techniques
- Tools
- Awareness issues
- Workflows
- Maturity levels

With this, the issues that surround both digital preservation and digital forensics are discovered as well as the benefits of adopting digital forensic tools and methods into a digital preservation workflow. The capabilities of digital forensics are explored as are the risks and complications. Digital forensic tools in use within collection institutions that are performing at a higher maturity level of digital preservation are explored. Workflows are investigated as they were revealed to be a major issue within collection institutions and determined to be the best way to provide the proposed enhanced solutions.

Chapter 3 METHODOLOGY begins by discussing how the public data were gathered and analysed. Based on these data, the questionnaire was developed which allowed specific and targeted data collection to be conducted via participating collection institutions. The process for developing the questionnaire is followed by the results and how they were evaluated.

The remainder of the methodology discusses how the digital forensic tools were investigated, how the data were analysed based on the output, and lastly, the workflow enhancements. How the benefits were evaluated is presented to conclude the chapter.

Following the methodology, Chapter 4 AUSTRALIAN LAW IMPLICATIONS evaluates the risk of sensitive data and the laws within Australia concerning these data.

The structure of this chapter is as follows:

The Privacy Act and Principles are evaluated and the extent of exemptions for collection institutions are addressed.

Sensitive data and identifying information are investigated to determine what sensitive data is and how it is defined under Australian law. Ingest scenarios have been created to stress the importance of sensitive data and how it can or may already be an issue residing in collection institutions.

The relevant laws are presented with a focus on areas such as defamation which can overrule exemptions in some cases. Examples of this are provided where the exemption to privacy law is not always enough to defend against a defamation claim.

The handling of Aboriginal and Torres Strait Islander material within a collection is addressed. The difference in legal and ethical considerations based on this material is explored, presenting a different perspective for consideration.

With an understanding of sensitive data and the role it must play within digital preservation, the next step involved investigation into the tools used within collection institutions.

Chapter 5 WORKFLOW TOOLS – DATA GATHERING investigates two sets of workflows from the BitCurator Consortium, made up of collection institutions within the United States of

America, in which the tools used within those workflows were identifiable. Workflow evaluation was undertaken on the BitCurator Consortium datasets and the OSSArcFlow dataset focussing on the workflows themselves rather that tool usage. All state and national libraries of Australia were reviewed based on their publicly available information such as policies, donation forms, workflows, and any other information pertaining to digital preservation. The public information was critiqued based on the difficulty to discover it and the quality regarding whether it provided enough depth and transparency.

The information that could not be gathered through public sources was gathered via a questionnaire (Appendix B - Questionnaire. The questions were developed based on the missing information that could not be derived from public sources. The results from both the Australian and U.S datasets were charted and compared, providing a list of tools and the frequency of their use among the institutions.

The data returned on the preservation and digital forensic tools used by the institutions investigated provided a starting point for exploration. The features and purpose of each tool was reviewed to better understand its position and potential within a preservation workflow. This provided pointers around what to look for when investigating a solution into the discovery and handling of sensitive data. The benefits of analysing data are explored in the respective chapter.

Chapter 6 DIGITAL FORENSICS – SENSITIVE DATA involves the usage of digital forensic tools on real data derived from media donated to the university that has been used by a family of users. This replicates instances where a backlog of media resides in a collection without any recorded information regarding its provenance or donation.

The tools that were investigated have been evaluated and their output potential is presented. Discoveries that provide useful and contextual information to collection institutions are discussed as to how they can benefit digital preservation in contrast to a digital forensic criminal investigation.

The digital forensic tools were approached with a digital preservation perspective. Output examples are provided to show the extent of information discovery from hidden and obscure locations which can strengthen a collection or mitigate the risk of sensitive data being mishandled.

Figures are presented showing the different views and visualisations these digital forensic tools generate to further emphasise the benefit these tools can have within collection institutions. The figures presented in this chapter also provide a summary of the software features, with emphasis on the output and the various ways in which the software can be useful to a collection.

Chapter 7 WORKFLOWS evaluates existing workflows based on their design, notation, and structure. These have been evaluated to present various design options for collection institutions that do not have a dedicated workflow.

The U.S workflows were evaluated based on their donor agreements, the ability to capture sensitive data, and how sensitive data were handled once discovered. The Australian institutions could not provide dedicated workflow diagrams, but were able to provide descriptive steps and made some attempts to visualise their process. Some institutions provided screenshots of the steps in their process. Workflows were conceptualised based on correlations of all the data provided in the questionnaire.

The proposed solutions to the issues presented throughout this thesis are presented in newly created workflows which are free of design restrictions, allowing institutions to design and implement at their discretion. These solutions aim to cover a wide range of areas in which sensitive data may be an issue. The workflows provided have been modularised to avoid large and complex diagrams. Sub-diagrams are created to further remove complexity and to provide a user-friendly, easy to follow example.

The discussion in Chapter 8 DISCUSSION provides a review of each chapter and how the solutions may be implemented into a typical digital preservation use case.

Conclusions and recommendations presented in Chapter 9 CONCLUSION and RECOMMENDATIONS follow the discussion and finally future work and the efforts needed to further improve digital preservation processes with digital forensic tools and methods are presented.

# 2 LITERATURE REVIEW

## 2.1 Digital Preservation and Digital Forensics: A Comparison

There are vast amounts of literature present today for both digital preservation and digital forensics. Each field contains new pressing matters being addressed by their respective communities. Current research for both fields is still, however, not acknowledging the other to the extent possible. There have been many attempts in comparing the two fields, their methods, techniques, tools, and workflows, often from a preservation perspective looking to adopt forensics. It is apparent that there are many similarities and that some goals are closely aligned.

> *"the handling of data within digital forensics is centred around preservation aims"* (Kim and Ross, 2012)

Kirschenbaum et al., (2010) published a report with the purpose of introducing the field of digital forensics to those partaking in digital preservation, exploring the points of convergence between the two fields. It was their desire to increase contact between the experts of each field to help create opportunities for experience and knowledge to be shared. Kirschenbaum et al., understood the limitations of collection institutions, but still acknowledge the potential of digital forensic methods and techniques, offering a distinction between tools and procedures. In the event an institution cannot afford digital forensic technology, there is still much to gain from forensic methodology.

Kam Woods of BitCurator states:

> *"Digital forensics commonly refers to the process of recovering, analyzing, and reporting on data found on digital devices. The term is rooted in law enforcement and corporate security practices: tools and practices designed to identify items of interest (e.g. deleted files, web search histories, or emails) in a collection of data in order to support a specific position in a civic or criminal court case, to pinpoint a security breach, or to identify other kinds of suspected misconduct."* (Lazorchak, 2015).

Although the goals specified differ when applying these methods to preservation, there are many parallels in the process as, Woods suggests, such as: chain of custody, provenance, and storing data in a means that prevents damage or loss. This is further supported in the report written by John, (2012) "Digital Forensics and Preservation" from the Digital Preservation

Coalition (DPC). In the report the similarities both fields share are discussed and it further highlights that many repositories have turned to digital forensics as it offers solutions for issues such as effective curation, automation, management, and analysis. John further discusses the parallels between digital forensics and digital preservation lifecycles, stating digital forensic workflows place more emphasis on the preparation and conduct, whilst preservation is focused on long-term preservation and reuse.

> *"Appraisal and selection of evidence, data and records, and the maintenance of provenance, a chain of custody, are prominent in both fields."* (John, 2012)

Some of the key differences between digital preservation and digital forensics include the lifespan of the data which they investigate or analyse. The digital forensic data an analyst works with are closely tied to the case on which they are working. Once the case has been resolved, it is unlikely the data will used again (Dietrich and Adelstein, 2015). The data will typically be retained offline and without any on-going maintenance which is where digital preservation differs. Digital preservation's main goal is to preserve data over long periods of time, while ensuring the data are maintained, unchanged, and can be accessed in a meaningful way. Rowell and Potvin, (2015) also recognise the fundamental needs of memory institutions that are not typically addressed by the digital forensics community, such as: incorporation of ingest in workflows, collection management, and the provision of public access to the preserved data. Another key difference is that digital preservation will likely be conducted on data created on legacy hardware, whereas digital forensics is likely to be applied to modern devices such as current models of personal computers and mobile devices.

> *"In a world where technology changes so rapidly and everyone is looking for the next flashy app or visualization it's hard to advocate for a process that has the sole aim of keeping things exactly the same."* (Schroffel et al., 2018)

Further differences discussed by Dietrich and Adelstein, (2015) include the documentation created during the acquisition and analysis processes. Traditional forensic investigations involve the creation of documentation that is mainly used internally with no intention of providing context to external users. Digital preservation involves greater consideration regarding documentation, detailing processes for internal use, and documenting the metadata for external use.

One of the similarities that both fields share is the increase in demand. As time progresses, so does the use of digital technology and the purpose for which it is used. Harvey and Mahard, (2013) address the profound effect technology have had on the digital preservation landscape from the start of the twenty-first century, identifying the changes in approach towards “longevity, choice, quality, integrity, and access”. The focus of collection institutions has had to shift and accommodate new policies and procedures due to technological change. Forensic analysists, once focused on physical evidence such as DNA and explosive residue, have also had to shift their focus as criminal activity and methods have evolved with technology, leading to digital forensics (Harvey and Mahard, 2013).

Therefore, the use of technology in criminal activities was inevitably going to influence the requirements of digital forensics, increasing the need and demand for such an approach. Faster and more invasive tools are then required to keep up with this demand. As for digital preservation, increases in demand will occur as more of our history is discovered and as more born-digital content is created. Many institutions may be in their infancy regarding their preservation needs, but growth is inevitable.

Gallinger et al., (2017) conducted a study based on The National Digital Stewardship Alliance (NDSA) storage survey results of the major US memory institutions from 2011 to 2013. This study revealed the total digital content stored in collection institutions and how growth exceeded expectations. Institutions, on average, almost doubled their anticipated growth rates within two years. However, it is stated that there are outliers that skew the total averages, and they are not an accurate representation of individual institutions. The results reveal over 25% are within the highest quartile containing the largest storage, with equally as many falling in the lowest quartile. The median storage was 25 Terabytes (TB), whereas the mean was 1014 TB, evidently skewed by outliers. Although there are uncertainties and inconsistencies in the averages, growth is still a consistent variable. Many institutions underestimated their rate of growth, with the average meeting their three-year expectations within two years (Gallinger et al., 2017).

The NDSA survey continued and in 2019 it was revealed that approximately 57% of participants required more than 100 TB of preservation storage, up from 33% and 34% from the 2011 and 2013 surveys (NDSA Storage Infrastructure Survey Working Group, 2020). The anticipated growth indicated almost a third of the participants were in the highest quartile of expected increase in storage requirements. Expected storage requirements of 100-999 TB for the 2011, 2013, and 2019 participants were: 32%, 28%, and 38%, respectively. Storage requirements over 1 Petabyte (PB) ranged from: 18%, 16%, and 30%.

A recent survey study by the Open Preservation Foundation (OPF) provides new results that further emphasise storage increase and growth (OPF, 2020).

The participants of the OPF survey were made up of 98 institutions across Europe (51%), North America (35%), South America (3%), Africa (5%), Asia (3%), and Australasia (3%). Academic and research libraries made up 30.6% of the participants, national libraries totalled 12.7%, national archives made up 7.5% with remaining GLAMS ranging between 6.6% to 1.5%.

The results reveal that 75% of its participants have under 1 Petabyte (PB) of storage, with 12.1% containing 1 to 3 PB of storage, 4% had 3 to 5 PB of storage, and 8% had more than 5 PB of storage. Furthermore, 33% of these institutions expect a growth between 1-10% within 12 months. Thirty-three percent (33%) more expect growth between 11-25%. Ten percent (10%) expect growth of 26-50% and 11% of the participants expect 76-100% growth. There were a small number of institutions (2%) that expected between 51-75% growth, and the remaining indicated they were unsure.

The last decade has shown considerable growth in storage requirements and anticipated storage increase. Recent survey results indicate within the next three years, more collection institutions will contain Petabytes of preservation storage (NDSA Storage Infrastructure Survey Working Group, 2020; OPF, 2020). In the years to come, growth will result in exponential increase. If an institution experiences 100% growth on a 25 TB storage, this results in 50 TB of total storage. Whereas a growth is exponentially larger when dealing with PB of storage instead of TB, where 10% growth (100 TB) is still considerable in size compared to total storage sizes under 1 PB.

It is said that 90% of the world’s data has been created in the previous two years, and this is a long-running trend over multiple years (Loechner, 2016; Marr, 2018; SINTEF, 2013). Whilst much of the data are produced by social media and mobile devices, not currently on the agenda of all collection institutions, this growth in data will have a greater impact when tweets, for example, are more widely considered assets to be preserved. The British Museum has in fact been preserving tweets since 2013 (Meikle, 2013). The UK Government Web Archive preserves central government information published on the web. The web archive includes videos, images, websites, and tweets dating from 1996 to present (The National Archives, 2020). Australian archives regard social media management an essential part of the preservation strategy, as seen with the strategies published in the National Archives of Australia and NSW State Archives & Records, for example (NAA, n.d.; State Archives & Records, 2015).

One can safely assume, growth is inevitable and will result in exponential increases in storage requirements. With this growth, the need and demand for digital preservation will grow with it. Additionally, the more data that need managing and processing, the more beneficial digital forensic tools and methods become. The need for such tools and methods is already increasing as indicated by the OPF survey results where 25% of the institutions were actively partaking in digital forensics and approximately 40% were either developing the capacity for digital forensics or researching its viability (OPF, 2020). This also means approximately 35% of institutions were not involved with digital forensics.

With continuous growth, there lies another issue. As the intake of born-digital material increases, more strain will be added for each institution which may have an already existing backlog of physical materials in need of processing and digitisation. Greene and Meissner, (2005) conducted a thorough review of traditional archival processing which highlights and challenges many of the ideals and assumptions archivists are making regarding the importance of certain processing activities. If collection institutions are dealing with such issues, the addition of born-digital backlogs will further exacerbate the strain they feel. Therefore, it is important to handle born-digital material appropriately, aided by digital forensic tools and methods which can help alleviate some of the strain. With physical material requiring a hands-on approach, the stages of born-digital preservation that can be processed automatically will free up time that can be spent on other processing tasks.

### 2.1.1 Ethics, Privacy, and Legal

Sensitive data have been largely overlooked as archivists have been preoccupied with the technical issues of preservation (Moss and Gollins, 2017). Whilst ethics, privacy, and legal issues are considered, the extent of their consideration is not enough.

Given the capabilities of digital forensic tools and their ability to discover data that a donor may not wish to be found gives rise to new ethical and legal dilemmas (Lazorchak, 2015). Although this is the primary goal for a digital forensic analyst, using the digital material to support a claim in a criminal investigation, it can be a burden for digital preservation due the nature of preserving everything whilst avoiding making assumptions (Dietrich and Adelstein, 2015). Media acquired from raw digital sources is said to…

> *"often contain significant amounts of contextual information along with potentially private and sensitive information in both created content and file and system metadata. Identification and management of this supporting information can be critical to ensure compliance with donor or submission agreements"* (Woods and Lee, 2012)

> *"Electronic records often contain personal identifiers, discussions of sensitive subjects, or other information that may be subject to restriction or redaction."* (Lee, 2018)

Larson, (2020) addresses the issues that also lie in big data where personal and sensitive information can be derived from large, individually anonymised data sets, through the process of deductive disclosure. Deductive disclosure is a process whereby personal information is accessible through the combination of large, individually anonymised data sets, because the aggregation of that data generates connections between data that make individuals identifiable. The risk of this disclosure is difficult to assess given the large volume and complexity of big data. In the differential privacy project from Harvard University, "linkage attacks" were discussed where data such as gender, date of birth, and zip code were enough to identify most Americans. Using an anonymised healthcare database linked to these attributes, the health records of the Governor of Massachusetts were identifiable (Harvard University, 2011).

Digital preservation is therefore subject to ethical and legal implications in all situations.

Emory University, whilst working with the Salman Rushdie Papers identified a key ethical issue (Kirschenbaum et al., 2010). The issues lie in having access to personal data that may reveal behaviour that was never meant to be disclosed, such as online activities, medical history, and financial information. With born-digital data, the discovery of sensitive information is magnified and it may not be explicit information, but implicit inferences by correlating pieces of data together, identifying patterns, a common practice in forensic investigations (Moss and Gollins, 2017).

Vinh-Doyle, (2017) faced several ethical dilemmas regarding the balance between privacy and information value. This involved the appraisal of electronic messages (email) from government agencies where the users assumed a greater level of privacy, leading to open dialogue that would not normally occur. Controversial and politically sensitive information was discovered via the appraisal. Ving-Doyle's study suggests this to be a growing issue as the management of email has worsened in the last two decades, with many government employees managing their email in an "ad hoc fashion", despite enforced management policies. This issue is expected to grow exponentially (Vinh-Doyle, 2017).

Having access to this information could lead to malicious action such as blackmail, identity theft, and other harmful acts. It then falls on the person responsible to make the appropriate decision, which may not always be easy for multiple reasons. Should the truth be disclosed even if it means the person of interest may have degrading information revealed about them? Is it ethical to possibly degrade that person or is it unethical to withhold the truth? Whilst this information can be critical to a digital forensic analyst working a case, it is troublesome for digital preservation.

This must be, and often is, addressed in a donor agreement as it must stipulate what must happen in the event of such data being found. The agreement must further clarify the expectations about the management and access to the donated materials as well as any other conditions the donor may require (Kirschenbaum et al., 2010; Lazorchak, 2015). Redwine et al., (2013) published an extensive guide for donors, dealers, and archival repositories, providing a recommended checklist that covers the pre-acquisition to post-acquisition stages of preserving donated material. The checklist ensures that various possible outcomes are considered, and that the donor and the collection institution capture as much information as possible and avoid any accidental changes to the data. This includes, but is not limited to:

- Information gathering and surveying
- Communication
- Privacy and intellectual property
- Legally protected files
- Agreements
- Transferring of materials

The acknowledgement of sensitive data, the volatile nature of metadata, and the need to involve donors provides a solid foundation on which to build and start a preservation workflow effectively.

Sensitive data can be highly valuable, but each case will be unique. Without a donor agreement, any decisions made will have some form of consequence. There may still be consequences involved if an agreement was met and the donor did not fully understand parts of their agreement in the event certain clauses were not explained carefully enough. With a detailed agreement, the risks involved with the decision-making are somewhat mitigated. This is particularly important when dealing with an individual who is deceased and can no longer have a say in what happens with their data. This also applies if the donor is no longer available or has revoked all claim on the data. In this case, a morally grey area is entered where there is no obligation to hide any data, but by revealing everything, a tainted image could be created of the original owner of the data. Context and relativeness must be questioned in this instance in order to decide whether this information should remain inaccessible to the public.

Context is an important aspect, especially in the digital age where information, such as video footage or quotes, taken out of context, can drastically change the nature of that information (Guerra et al., 2017). The digital makeup of this information makes it easy to change context intentionally. The publication date of information may also be a significant factor of context, as the example by Moss and Gollins, (2017) describes:

> *"consider a text from the seventeenth century that describes religious or ethnic minorities. Now consider if that identical text were to be authored and published today. In the context of a historical document, the language (even though now reprehensible) would generally not be considered particularly sensitive. In the context of a modern document, the reverse would be true. From this we can see that the zeitgeist of publication (the context in which it was said) is also critical."* (Moss and Gollins, 2017)

Furthermore, cloud computing ushered in a new era of issues for both digital forensics and digital preservation. The reason for this being, regarding legality, that data stored in the cloud are often held in multiple locations which may span across different jurisdictions (Narayana Samy et al., 2018; Quick and Choo, 2013; Rahman and Choo, 2015). In the United Kingdom, Scotland's codification of sensitivity is distinct from the other member countries (Moss and Gollins, 2017). Therefore, one can imagine the myriad of jurisdictional differences worldwide.

This is more problematic for criminal investigations, given that they are time-sensitive and cooperation is required with the local law enforcement of the datacentre's jurisdiction in order to legally gain access to cloud stored data (Hofman et al., 2017; Moss and Gollins, 2017). Issues regarding privacy, trustworthiness, reliability, and jurisdictional differences are also explored by Hofman et al., (2017) in which the suggested solution involves revisiting privacy models and reframing them at a record level.

> *"Acquisitions that cross national boundaries can be challenging because of the different copyright and intellectual property laws and practices around the world."* (Redwine et al., 2013)

Cloud computing is not the only online aspect in need of consideration. Online catalogues place collection institutions in the centre of the ongoing privacy debate…

> *"When records are transferred to an archive there is a clear expectation that they will be made public. Digitally born records come with the same expectation, such as in the current plans of the National Archives of the United Kingdom. Once online the content will be indexed by ubiquitous web search engines and content will be easily discoverable in a way it was not in the analogue world. This places the archive at the centre of this privacy debate, whether most archivists have realized this or not."* (Moss and Gollins, 2017)

Information regarding the specifics of the legal and ethical issues for collection institutions within Australia is explored in Chapter 4 AUSTRALIAN LAW IMPLICATIONS, which focuses on legal documentation and existing laws. Legal and ethical matters surrounding indigenous data belonging to the Aboriginal and Torres Strait Islander people are addressed.

Although these matters surrounding Aboriginal and Torres Strait Islander people are specific to them, this should also be a global consideration. The Global Indigenous Data Alliance (GIDA) have published a one-page document on the "CARE" principles for Indigenous Data Governance (GIDA, 2019). "CARE" stands for: (Collective Benefit, Authority to Control, Responsibility, and Ethics).

> *"The UN Declaration on the Rights of Indigenous Peoples (UNDRIP) reaffirms Indigenous rights to self-governance and authority to control their Indigenous cultural heritage embedded in their languages, knowledge, practices, technologies, natural resources, and territories (i.e., Indigenous data). Indigenous data, which include data collected by governments and institutions about Indigenous Peoples and their*

> *territories, are intrinsic to Indigenous Peoples' capacity and capability to realise their human rights and responsibilities to all of creation."* (GIDA, 2019)

## 2.2 Provenance

Digital forensic tools and methods can advance the three fundamentals of digital archival practices which are: provenance, original order, and chain of custody (Lee, 2012, 2018). Lee describes the three fundamentals as follows:

The provenance in archival context often identifies the origin or source of a record, however, it also captures the inscription, transmission, contextualisation, and interpretation. This accounts for the existence of the digital material, as well as the characteristics and continuing history (Lee, 2012, 2018).

The original order principles indicate digital materials should be stored and organised in a way that reflects their original creation environment. The original order can often reveal information about recordkeeping, and it facilitates navigation and access (Lee, 2012, 2018).

The chain of custody is a record of all those who have held digital materials from the moment they are created through to ingest. This information is important for legal compliance, authenticity, evidential integrity, and legal admissibility. Alongside data, documentation and records should also be kept of the state of the material and any changes for each custodian (Lee, 2012, 2018).

Provenance can only be achieved with full transparency regarding data and its history.

> *"Trustworthy data and records, that is, data and records that can be presumed authentic, reliable and accurate, and are useable and readable, rely on intellectual controls, protective measures, data partitioning and processing, legal compliance and risk management, identity and access management, service integrity, and endpoint integrity, that is, on factors that can be "objectively" assessed in order to establish trust in records and data online inferentially. Increasingly, we also want to be able to know the original source, provenance, and chain of custody of the records and data. Thus, in the words of Weinberger, "transparency is the new objectivity""* (Duranti and Rogers, 2016)

Transparency is one of the seven core preservation attributes proposed by Kim and Ross, (2012), aligned with provenance in that:

> *"Any tools and specifications involved in the format should be a publicly published open standard and non-proprietary to avoid restrictions regarding activities that support long-term preservation and access of material in the archive, such as making modifications to the format, distributing new versions, and tracing accountability and authenticity."* (Kim and Ross, 2012)

Provenance is therefore crucial to both digital forensics and digital preservation. For digital forensics it can prove the origin of digital material, the ownership, and it can prove if some form of manipulation or tampering has occurred. Raghavan and Raghavan, (2014) analyse the use of metadata association modelling and the use of metadata to determine digital image relationships which can be used to identify doctored images and instances of intellectual property theft. The approach is not without challenges, one of which is false-positives, where unconnected files may be associated, and more likely to occur in shared environments containing multiple computers.

This is equally as important for preservation as provenance serves as a means for quality assurance which is essential given that a large amount of data comes from donations and the web or cloud where it is subject to replication, query processing, modification, and merging (Hartig and Zhao, 2010; Raghavan and Raghavan, 2014).

Unless donations are delivered directly to the collection institution by the creator, intermediaries will be involved, complicating and jeopardising trustworthiness by mere acts of opening files or booting up a computer (Kirschenbaum et al., 2010). The greatest concern is not with format obsolescence, emulation or migration, but in the inconsistent use of systems by people where these variations significantly impact the preservation workflow (Moss and Gollins, 2017). Moss and Gollins further state that most of the difficulties are not the preservation aspects of the workflow, but from other archival challenges, specifically describing and presenting material for use.

Whilst the provenance is necessary to prove the integrity of digital material, the provenance itself must also be trustworthy. Provenance metadata like any other metadata is potentially at risk of removal or modification.

> *"An authentic source may be deceptive or unreliable, and although reliability is an important component of trustworthiness, the veracity of a document's content is often not the concern of archivists working with cultural heritage materials. Rather, the provenance of both analog and digital materials, as well as documentation about their*

> *storage environment, what has been done to them, and by whom, are the key aspects of establishing and maintaining trust."* (Kirschenbaum et al., 2010)

This is increasingly likely when dealing with data extracted from the web, especially cloud environments. In an ideal situation, completeness, integrity, availability, and confidentiality must be guaranteed for the provenance, which in turn ensures the same level of guarantee for the material in question (Cho and Chen, 2018; Hasan et al., 2009). Completeness ensures all the records are present and no key data are missing. Integrity can be ensured if no alterations or forgeries have been or could be made, ideally. Availability allows auditors to verify the integrity and confidentially ensures this access is strictly for authorised users only.

There are many views on provenance as are there many models. In the model shown in Figure 2, provenance is viewed from different perspectives. For example, as stated by Zhao and Hartig, (2012), referring to the W3C PROV Model Primer (W3C, 2013), three perspectives are described in which provenance can be viewed; Agent-oriented, object-oriented, and process-oriented. Agent-oriented describes information about the entities responsible for generating or manipulating information. Object-oriented traces the entities that contribute to the existence of other entities. Process-oriented tracks the actions and steps involved in generating or manipulating information (Zhao and Hartig, 2012; Zhao et al., 2016). Combined, these perspectives capture the 'who', 'what', 'when', and 'how' which can be seen in Figure 2.

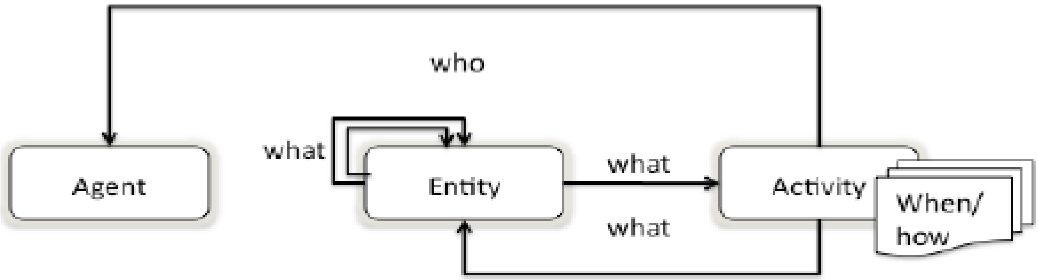


Figure 2 - Provenance Described from Three Perspectives (Zhao and Hartig, 2012)

This is what is known as a process-centric modelling pattern (Zhao and Hartig, 2012). An activity is always introduced to describe creation and modification, which in turn describes the relationship between entity and agent; with the exception of entity to entity relationships where an activity does not have to be introduced (Zhao and Hartig, 2012).

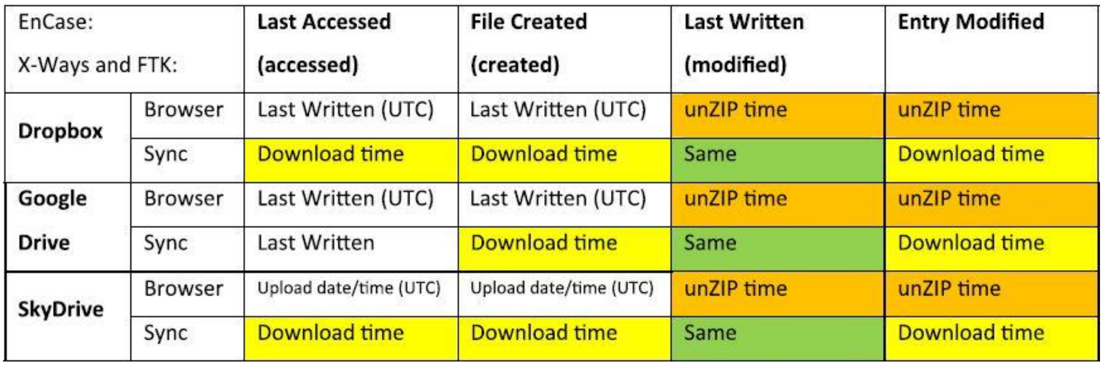
The way to identify and describe provenance is through metadata (Raghavan, 2013). Many of the digital forensic methods and tools primarily work with metadata, however, they often have a forensic perspective with specific objectives. Some metadata types a forensic analyst may be concerned with include: File name, File Extension, File Size, Hash Value, Date Last Accessed, Created, and Modified. The types of values may include: application, document, file system, email, embedded, business, and geographical metadata (Raghavan, 2013).

Digital preservation also has use for such metadata. Whilst technical and system metadata may be adequate to support the basic preservation of digital artefacts, what these metadata do not capture is the context of the material being preserved or analysed. This is important for the preservation field as meaningful access to preserved materials is a primary goal in contrast to being supplied as evidence in a criminal investigation. Maintaining availability, identity, persistence, renderability, understandability, and authenticity is the primary objective of preservation metadata, which encompasses various different types of metadata (Gartner and Lavoie, 2013). With this, the FAIR data principles are encompassed, that is to be: findable, accessible, interoperable, and re-usable (Wilkinson et al., 2016).

Having accurate and complete provenance metadata of digital material is important, especially metadata that indicate how the material was originally used and accessed. This may range from how the material is viewed, how it is handled, or how it is heard. Modern technology will not provide an authentic look and feel for legacy artefacts, but emulation provides the best means to accurately achieve an authentic interaction and viewing experience (Cochrane et al., 2018).

Even if an emulated environment is provided that closely replicates the original system platform, there must be awareness of how the material was intended to be handled. Today the most used mice are optical mice, however, the material in question may have originally been created using a different or unique peripheral device. For example, a current peripheral device may contain features that invite certain gestures and interactions that may change the overall experience for the user; the device may be missing features that allow interaction that was possible with an older device (Dietrich and Adelstein, 2015). Therefore, it is important to know as much as possible about the provenance of our digital material, which is why documentation and metadata are essential.

A matter of concern is the authenticity of files that have been processed in the cloud. Given the nature of how cloud-based storage works, the different types, and the different ways to access or use the cloud, the concern is justified. Studies by Quick and Choo, (2013) investigated what happens to files after they have been processed through the cloud, resulting in interesting discoveries. The files themselves did not change during the process of uploading and downloading using Dropbox, Microsoft SkyDrive, and Google Drive. This was determined by using MD5 (Rivest, 1992) and SHA1 (Eastlake 3rd and Jones, 2001) hash values. The timestamp information, however, varied. The “last written” (modified) time was unchanged when downloading files using a client, however, this was not the case when using the browser-based functions. Figure 3 shows the different effects that occurred on the timestamps for each cloud provider tested.



| EnCase:<br>X-Ways and FTK: | | **Last Accessed**<br>**(accessed)** | **File Created**<br>**(created)** | **Last Written**<br>**(modified)** | **Entry Modified** |
|---|---|---|---|---|---|
| **Dropbox** | Browser | Last Written (UTC) | Last Written (UTC) | unZIP time | unZIP time |
| | Sync | Download time | Download time | Same | Download time |
| **Google Drive** | Browser | Last Written (UTC) | Last Written (UTC) | unZIP time | unZIP time |
| | Sync | Last Written | Download time | Same | Download time |
| **SkyDrive** | Browser | Upload date/time (UTC) | Upload date/time (UTC) | unZIP time | unZIP time |
| | Sync | Download time | Download time | Same | Download time |

Figure 3 - Timestamp Changes (Quick and Choo, 2013)

Three of the four metadata elements tested were affected except for the one instance where the client sync did not change the “Last Written” time for Google Drive. This issue aside, there are further pressing matters relating to the cloud computing. Some of these issues are discussed by Roussev and McCulley, (2016) in which they describe how one cannot assume the client is the original source of the data as it may only hold a cached, incomplete, or out of date version of the original data. Some cloud providers offer selective replication, meaning only select data are held in storage on the client’s side, allowing devices with low storage such as mobile phones to better utilise cloud services (Roussev and McCulley, 2016). Software as a Service (SaaS) such as Google Docs is an example of where the documents are stored as only a hyperlink on the client-side (local disk).

The impact of data changes may be lessened if the changes can be accurately tracked through the change history (provenance), however, the cloud makes this challenging. The cloud is a different environment that is highly scalable, not extensible, and the introduction of latency (the time it takes data to travel from client to server) presents different update and error models, for which existing provenance systems are not designed (Awad et al., 2016; Muniswamy-Reddy et al., 2010). This is further supported by Duranti, (2016) where she presents the risks associated with cloud computing from an archival perspective. The risks to metadata, transparency, and security are discussed, as are the issues with utilising a preservation cloud service, which is the suggested solution to the issues discussed. Giving control to the cloud means losing control in the archive. This also aligns with the risk of cloud provider reliability.

Cloud computing is not the only online risk to provenance and trustworthy data; all online environments pose risks. The "Internet of Things" which allows interactions between connected devices and people results in a loss of transparency as well as increased security and privacy issues (Duranti and Rogers, 2016). The more services that are reliant on an online environment, the more fragmentation of information will occur and the control over this information will diminish (Duranti and Rogers, 2016).

Many other challenges hinder both digital preservation and digital forensics, however, the impact may be greater for a forensic analyst who may be investigating a time-sensitive and priority case. If issues arise preventing the preservation process from proceeding, there may be alternative solutions to explore with more time to do so. Many of the issues and threats discussed throughout this study may be rare situations, but the probability of such incidences happening are always increasing the more digitised the world becomes (Duranti and Rogers, 2016). Volumes of data and the complexities of data are increasing; the way data are stored and handled is changing, all of which present difficulties for both digital preservation and digital forensics.

## 2.3 Digital Forensic Methods and Techniques

Digital forensics offers valuable methods and techniques that advance many of the goals within collection institutions such as maintaining authenticity, describing records through metadata, and providing responsible access (Lee, 2012). As there is a substantial amount of information residing in the underlying structures of computer systems, all archival functions can benefit from the computer-assisted appraisal and selection methods digital forensics can offer (Lee, 2018).

> *"Every step of the archival lifecycle may be influenced by the forensic approach. Even the integration of digital with analogue is embraced by the forensic workflow, with digitized objects being imported into the digital forensic case – accordingly, Forensic Toolkit (FTK) has an OCR (Optical Character Recognition) capability."* (John, 2012)

There are many forensic methods and techniques in use today within memory institutions, however, there are far more advanced techniques not readily known  (Lee, 2012, 2018). Some of the more common methods and techniques include preventing unintentional and irreversible changes to source media with the implementation of write blockers, including the prevention of changes to timestamps when copying disk contents to another storage device (Lee, 2012).

Creating disk images can mitigate potential hardware failure and allow further extraction and analysis tasks to be performed (Lazorchak, 2015; Lee, 2012; Woods and Lee, 2012). Digital forensic disk images contain embedded capture metadata and redundancy checks which detail a technical capture record which can improve the survivability of raw images in case of storage failure. This is important when considering long term storage (Lazorchak, 2015). When data are imaged, cryptographically secured hashes are created and are used to compare with the data after analysis to ensure no changes have been made, further ensuring the integrity and validity of the data (Dietrich and Adelstein, 2015). Hashes can be generated using MD5 which is quick, and most of the time reliable; however, there are cases where two files, although different, have generated the same value. SHA-1 is another hash generator, more complex and reliable, however, its generation time is longer than that of MD5. There are stronger versions of the SHA algorithm, such as SHA-256 (Nohe, 2018).

It should be noted, however, that many collection institutions choose to accept materials as files rather than disk images, possibly due to technical barriers and the assumption that they are not always suitable for certain types of materials (Wiedeman, 2016; Woods and Lee, 2012). However, with the use of write blockers and disk images, it is easier, more efficient, and reliable to ensure provenance, original order, and chain of custody; as well as enabling further digital forensic processes (Lee, 2012).

Another common method is the use of hex editors (Dietrich et al., 2016). When standard tools are not able to provide bit-level data, which may not be easily identifiable, making use of hex editors allows the user to analyse the hexadecimal bit stream representation of a file. This can provide clues about the file, even if the file is corrupted and can no longer be accessed via conventional means (Dietrich et al., 2016).

Metadata discoverability is a core function of many digital forensic tool and methods. Methods such as these are useful for various tasks and can benefit many disciplines. If a photographer is assessing a digital photograph and wishes to understand how the image was captured, this information can be derived from metadata and may reveal the hardware used to capture the photograph and settings used (Hart, 2015; Hart and de Vries, 2017). However, these data are not always available. The tests conducted by Hart, (2015) and Hart and de

Vries, (2017) illustrate how metadata can be removed from a file during transit, such as being uploaded to social media websites.

Furthermore, there are methods for establishing the exact camera a digital image was taken from. One method in particular is known as Sensor Pattern Noise proposed by Lukas et al., (2006) and verified by Khanna et al., (2009) in the FBI archives. This method involves a reference pattern being estimated for each camera to be used as an identifier, much like a fingerprint. Noise is extracted from the digital image and then correlated against a collection of camera patterns. This method has been shown to produce results close to 100%. This method is primarily used in forensic cases against seized evidence, allowing the analysts to prove ownership of certain images by correlating them against cameras found in the person of interest’s possession. Such methods set an example of how digital forensic tools and methods can provide enhanced data gathering, useful for collection institutions aiming for complete and accurate metadata for their items.

Digital forensic tools can further automate the triage process, prioritising items that are targets for preservation or that require attention. This reduces the need for manual triage which reduces the requirements of trained professionals making complex decisions (Lazorchak, 2015). However, human intervention is still necessary for automated processes. Quantifying the success and failures of digital forensic tools, identifying false-positives and false-negatives, still requires the judgement of qualified professionals (Chassanoff et al., 2016). The Stanford University project to capture and process born-digital files for the STOP AIDS project met this challenge of time-consuming human intervention required with pattern searches due to false-positives (Wilsey et al., 2013).

Triage is one of the various methods proposed to address the increase in data volume, the others include: data mining, data reduction, parallel or distributed processing, artificial intelligence, and digital forensics as a service (Quick and Choo, 2016). Quick and Choo present a method of data reduction called the Digital Forensic Data Reduction by Selective Imaging, suggested to be run alongside a typical forensic process, allowing rapid triage, collection, review, and archive forensic data to support the forensic process.

During ingest, the redaction of sensitive and confidential information may be required as complete disk images of computer systems are often processed.

> *“an increasing number of personal data collections, in the form of digital media and complete computer systems, are being offered to the academic institutional archive”* (Knight, 2012)

This is where information such as phone numbers, email, addresses, and other sensitive data may be present (Dietrich and Adelstein, 2015). This is a key workflow step that must be conducted before access is granted to the collection (Meister and Chassanoff, 2014). Digital forensic tools enable methods that allow the searching of private and sensitive data, with a choice of other items of interest. This is most useful when redaction of sensitive data is required (Lazorchak, 2015). This must be stipulated in the donor agreement, however, if there is no agreement, these data will be preserved and seen as potentially valuable. Sensitive information may also be useful when dealing with materials such as societal, cultural, and historic data that may be donated to collection institutions without any supporting documentation (Woods and Lee, 2012). This is where having a disk image is advantageous as it provides environmental context which aids in finding likely locations for passwords, encryption keys, generated wordlists for password recovery, and in some cases, circumvent the protection (Woods and Lee, 2012).

Furthermore, with a hard disk or disk image, significantly more data can be extrapolated pertaining to a featured collection piece, providing more context on how it was created or how it is meant to be used or examined.

> *“would not the documentation that is gathered, made, and collected communicate more about a work, and how it is experienced, than its physical manifestation?”* (Dekker, 2018)

Bartliff et al., (2020) and their study into the late Stephen Dwoskin (1939-2012), an influential filmmaker at the forefront of the shift from analogue to digital film, explored the potential of digital forensic tools in this context. Their findings illustrate that the social, material, and temporal dimensions of digital forensic analysis also align with the “distributed creativity” notion, explored in (Glăveanu, 2014). With access to the media on which an artist’s content was created, substantial information can be discovered that give insight into the creative process, the environmental and technological context, and various other elements, such as influences and personal or professional history, that may factor into the artist and their work (Glăveanu, 2014). This notion was earlier explored by Gareth Knight, who stressed the benefits of digital forensic techniques allowing for greater breadth and scope when collecting information. Knight discusses the identification of “lost” material and being able to discover abandoned or previous versions of work which can provide contextual information into the user’s creative process (Knight, 2012).

With the benefits of digital forensic tools and methods clearly established, the goals the digital preservation community should focus on regarding the adoption and utilisation of digital forensics, according to Lazorchak, (2015), include:

- Better workflow modelling,
- sharing information and standardising vocabularies to describe the actions taken using digital forensic methods for preservation purposes,
- improved community driven documentation for digital forensic tools, but from a preservation perspective,
- and rather than relying on a tool tutorial to educate librarians and archivists, provide instructions on the process and link that to a tool.

With this, collection institutions that choose to adopt and utilise digital forensics, or may already be doing so, can document their procedures, success, and failures, so their peer institutions may have an example to follow for guidance. This is a step towards the standardisation of digital forensics within digital preservation communities.

## 2.4 Tools

Digital preservation and digital forensics each have limitations. When a forensic analyst is investigating a criminal case, there are time and legality issues, however, the restrictions regarding the hardware and software they can use is based on whether a tool has been verified and validated through testing, and if it has been approved by the courts (Vincze, 2016). Digital preservation, specifically performed in public collection institutions, may have resource, policy, and budget restrictions which will dictate their choice in hardware and software (Velte and Wikle, 2020).

Those undertaking digital preservation may not always be in an environment that allows full control over their systems, which may be the result of permission and access privileges. However, many of the tools developed for digital forensics and digital preservation purposes do not require that level of permissions and can often be run in virtual environments (Dietrich and Adelstein, 2015). This gives the preservation community an abundant source of potential tools to utilise should their budgets not allow the purchasing of proprietary software.

It has also been long argued that essential content, structure, and context elements for digital files can reside in multiple data sources, rather than a single file (Lee, 2012). Digital forensic tools and methods accommodate this and allow archivists to treat data at a lower level, bypassing filesystems and allow raw bitstreams to be read which can then be decomposed into appropriate records (Lee, 2012).

Preservation communities make use of many tools such as Duke Data Accessioner (Shaw, 2017) for migration, FITS (Harvard Library, 2018), fido (OPF, 2010), and DROID (The National Archives, 2018) for metadata extraction and validation (Dietrich and Adelstein, 2015).

The growth of data and information has exceeded the scope of manual maintenance, increasing the requirements of tools, especially those that automate metadata generation and extraction (Dobreva-McPherson et al., 2013). Dobreva-McPherson et al., (2013) describe workflows derived from their background research where material received into repositories is accompanied by metadata or the metadata are generated after ingest. The issue here is that in both scenarios, metadata quality control is lacking. The quality of metadata impacts discovery, retrieval, data and preservation management, and access. The solution proposed by Dobreva-McPherson et al., involves ensuring metadata quality control as a pre-ingest process, identifying the need to initialise a workflow correctly. With complete and accurate data, digital objects can be represented in collections as they should be and managed accordingly.

Dietrich and Adelstein make note of BitCurator (BitCurator, 2018), an environment that rather than developing everything from scratch, is made up of many existing tools and forensic utilities, adapted to meet the needs of archives and preservation whilst being mindful that not all users are experts. Most digital forensic tools are not designed with archival goals in mind, therefore, BitCurator recognises and attempts to bridge the gap between the original law enforcement context of digital forensic and the cultural heritage context (Lee, 2012; Rowell and Potvin, 2015).

> *“The BitCurator project is a joint effort – led by the School of Information and Library Science (SILS) at the University of North Carolina, Chapel Hill and Maryland Institute for Technology in the Humanities (MITH), and involving contributors from several other institutions – to develop a system for librarians and archivists to incorporate the functionality of many open-source digital forensics tools into their work practices”* (Lee, 2012).

The case study presented by Meister and Chassanoff, (2014) describes the results of using BitCurator to process a real world, born-digital, collection, in which all the workflow requirements, long-term preservation and access requirements, were able to be met.

Rowell and Potvin, (2015) summarise the main features of BitCurator that support the gap as: acquisition, reporting, redaction, and metadata export. Tools such as these and the techniques they allow, should they be properly incorporated into archival and preservation workflows, will greatly improve the archival procedures. Kirschenbaum et al., (2010) identify the benefits to include: being able to capture more information from the data, helping repositories manage data more efficiently and with standards, reinforcing documentation in all aspects of the curation cycle, and allowing users to preview the contents of their data. However, Kirschenbaum et al. emphasise the difference between tools and procedures, stating:

> *"Technology is expensive, but methodology is free"* (Kirschenbaum et al., 2010)

Not every institution can incorporate or adopt a complete digital forensics workflow as every institution will differ in goals and requirements. It may not always be viable to purchase a forensic workstation or software if the institution is only dealing with one type of material. By utilising the methodologies digital forensics has to offer in conjunction with open-source free solutions, the institutions goals can be met.

Whilst software-based tools are generally created for a dedicated purpose, they can often be repurposed for other needs. Both digital forensics and digital preservation process video files ranging in size and duration. A forensic analyst analyses video footage carefully for suspected material such as criminal evidence, terrorist propaganda, and other legal issues. With digital preservation, video is analysed for context, errors, fragmentation, and content in need of redaction before determining if the video should be added to their collection. Being able to perform this task efficiently and accurately is something that can be achieved with digital forensic tools and methods. Video thumbnail methods, as shown in Quick and Choo, (2016), considerably reduce the time taken to analyse video files. This allows frames to be selected at set intervals, for example every five seconds, and a thumbnail is generated each time, resulting in a digital image that can be analysed quickly. This also reduces the size of data being dealt with, which offers more suitable file transfer and storage. The example presented had a video file, 750 MB mp4, thumbnailed every 8-10 seconds, resulting in a 176 KB jpg file, a significant reduction. Further sampling was conducted, taking a 500 GB hard drive, 399.3 GB of which comprised 828 video files. The thumbnail conversion resulted in a total of 134 MB of snapshots, 0.0034% of the original size.

Standalone tools exist to perform this task, such as ThumbnailMe (Rousseau, 2012), used in the example in Quick and Choo, (2016). This technique can also be found in digital forensic software suites such as Autopsy (Basis Technology, 2018a), which allows a video triage plugin to be installed, which performs in a similar way.

There is no single solution that suits the needs of every institution, which is why complete solutions such as BitCurator and the alternative of standalone tools, both open-source and proprietary, each have their place. Open-source solutions are becoming the popular choice as costs rise in digital preservation. In the OPF, (2020) survey results, 68% of participants reported increases in digital preservation costs over the last 5 years with 76% of all participants predicting an increase in cost over the next 5 years. With this, 94% are using some form of open-source technology. This includes primarily standalone tools, with approximately 40% using open-source solutions embedded in commercial systems. Only 6% of participants use no open-source, with 30% of the institutions being entirely open-source.

There are many solutions that require investigation and it may take some trial and error to find the solution that fits as Bentley Historical Library discovered (Eckard and Hagen, 2018). Eckard and Hagen report on the revamp of Bentley Historical Library's removable media workflow where they describe the process of looking to adopt BitCurator. Ultimately, they found what was already in place was suitable and the archivists were not having trouble with the tools they had, they did however desire workflow improvements to mitigate issues with large-batch processing. One case presented is the use of the Archivematica (Artefactual, 2019) pre-transfer event tracking and how the existing workflow did not support this. Eckard and Hagen identify and acknowledge that the digital forensics approach adopted by their peer institutions could aid them in meeting their own needs.

Cross-institutional learning and the importance of transparency has been recorded throughout the history of digital preservation. These are influenced by accurate and concise workflow documentation. Since the 1990s, collaboration and transparency have been essential in meeting several key milestones (Baucom, 2019). Baucom emphasises that no digital preservation achievements from the past were developed in a vacuum and the future achievements will be built upon the past. Baucom further states that open and accessible information is key in helping collection institutions choose carefully how to use their increasingly limited resources and how they can improve their digital preservation workflows.

### 2.4.1 Tool and Solution Awareness

It is still apparent in current literature that the potential for preservation goals to be met using digital forensic tools has not been fully realised. For example, the following statement from Wiedeman:

> *"Forensics tools are designed for digital forensics investigations not archives. In most cases, tools choose to extract large bodies of contextual metadata from disk images, not—as we require—to gather metadata on individual files. This means that most forensic tools available today often take too long to gather record-events for individual files."* (Wiedeman, 2016)

This statement shows that the applicability of digital forensics to preservation is still not fully appreciated. Wiedeman stated that many digital forensic tools mainly extract large bodies of metadata from disk images and not individual files. However, in the BitCurator environment, for example, many of the tools used on disk images can be used on live filesystems. There are the options of using a disk image, a live disk, or the user can manually select a directory which in turn will significantly reduce processing time and allows focus on specific files. Many of the tools within BitCurator have standalone versions, therefore, the workload can be reduced if the user knows what they are looking for and where it may be located. With this information, specific tools can be selected and used on a specific group of files, reducing the need for large bodies of data being processed.

The different points of view are the source of other issues surrounding the tools available. For example, the tools developed for criminal investigations are done so with current technology in mind. The tools need to be updated to ensure they are compatible with current technology. The cases for which these tools are designed are likely to have a short lifespan in contrast to collection items. Therefore, support for post-processing activities may not be considered. Digital preservation is concerned with long-term goals in that access is to be an on-going process. The data preserved is likely to be retired from legacy media and devices. Therefore, utilising the tools used in digital forensics may require modification and adaption. Tools may also be discontinued once outdated as it becomes increasingly difficult to support older hardware and software (Dietrich and Adelstein, 2015). Whilst backwards compatibility and archives of older versions exist, there are limitations in their support and availability.

Examples of this have been listed by Microsoft regarding products ending support in 2021 (Microsoft, 2022). This hinders digital preservation as new methods must be adopted and are often required to create custom solutions, adding to time and complexity. Furthermore, a lot of the reverse engineering and forensic work conducted on current technology may be of use to the preservation community long after its usefulness for the digital forensics community (Dietrich and Adelstein, 2015). Therefore, sharing this information is very important. If the two communities were made aware of each other's needs and started working towards closing the gap, the software developers for these tools may also be influenced and start to address these issues. This will make the cross-discipline benefits easier to obtain.

One study on community-based digital preservation identifies the many approaches and solutions to choose from regarding preservation, some of which promote international collaboration (Trehub et al., 2018). This revealed over sixty tools and services were available in 2013 which by 2018 had grown considerably, but also included comprehensive "DP networks and turnkey solutions". The community owned digital preservation tool registry lists 554 different tools (COPTR, 2021). There are many tools and solutions, hence the lack of standards. With so many options it is hard to determine which solution is best fit for a specific institution. This is especially problematic when a group of similar collection institutions are all using different and various tools and are not transparent about their operations. The problem again lies in awareness and whether information of these solutions is available.

Another aspect of consideration regarding tools and awareness, is that the nature of records are fundamentally changing (Moss and Gollins, 2017). The example of this change provided by Moss and Gollins describes the significance of material originally intended to be used short-term. The example describes tweets from Donald Trump, and how a 280-character message on Twitter, originally intended as a quick, short-term, digital object, must now be considered to have long-term significance. This may change the nature of tools required. Perhaps social media integration and machine learning aspects will be more prominent in the future of collection institutions as these once-seen as insignificant digital objects are now requiring our attention.

Awareness may also be the result of training. The skills and capabilities developed through training can increase the overall knowledge of preservation and archival activities, allowing those partaking to be better equipped for finding solutions. Cunningham, (2008) investigated the curation and archiving experiences of the National Archives of Australia (NAA), identifying the challenges they face. Regarding the skills and capabilities required, Cunningham states, as early as the 1990s, it was stressed how much the archivists of that generation needed to know. This was in reaction to an influx of desktop computers, online networks, and other emerging technologies. Furthermore, in 1998 the NAA experimented with the Monash University training course. This was an early recognition that the current skills and training were not adequate for the challenges to come.

> *"While we can and must forge partnerships with other professions, such as information and communications technology professionals, lawyers, business analysts, communications experts, and educators, today every digital archivist needs a range of knowledge, skills, and qualities."* (Cunningham, 2008)

Following this comment, Cunningham provides an extensive list of the knowledge, skills, capabilities, and qualities needed by digital archivists. The list covers a wide range of areas, all of which better prepare someone for the acquisition and archival tasks associated with born-digital material and is still currently relevant.

In 2011, in the report "New roles for new times: Digital Curation for preservation", it was recommended that digital curation be seen as a core function and not focus all resources on the physical collection, using those resources to invest in long-term training programs, hiring of experts, and maintaining engagement with digital curation services (Walters and Skinner, 2011). In the same year, Michael Olson, digital collections project manager at Stanford University Libraries, started training library staff in forensic / logical capture and the use of FTK (Olson, 2011).

Furthermore, digital curator roles within collection institutions may not be utilised to their full potential, as Tammaro et al., (2017) suggest. In their research, participants from academic libraries, research centres, and data curation centres from Australia, Canada, and the United States were recruited and interviewed. Among the participants, various positions titles were held, including:

- Coordinator of data curation and scholarly communications
- Data curation librarian
- Data librarian
- Data curation scientist
- Digital curation coordinator
- E-research project officer

- Project scientist
- Research data management librarian
- Research services coordinator

It was discovered that all the participants were mainly responsible for outreach and training programs. Most of the participants discussed a mismatch in the perception of data curation responsibilities and the tasks they were performing. All participants stated they had not been part of any data management activities directly.

> *"data curation is more about providing information about good data curation practices to the people who need to curate their data or could be curating data"* – *Participant E* (Tammaro et al., 2017)

Whilst this was a preliminary study, a following publication in 2019 provided additional information (Tammaro et al., 2019). This included a larger participant list which expanded on the original three countries, adding: Austria, Germany, Netherlands, Sweden, Switzerland, and the United Kingdom. Tammaro et al., (2019) indicates that outreach and training is still the main responsibility of the participating curators, however, unlike the initial findings, a small number of participants were involved in technical services. Data management was still primarily consultative, offering advice on data management and planning. It was also emphasised that the curator role has become a support role to aid researchers.

To support this, Figure 4 displays the list of jobs that play a direct part in digital preservation activities derived from the 2019 – 2020 Open Preservation Foundation digital preservation community survey (OPF, 2020). This indicates that more than 60% of the surveyed institutions utilise the following staff in their digital preservation activities:

- Cataloguer or metadata analyst
- Director, manager, or administrator
- Digital archivist or curator
- System administrator
- Digital preservation officer or assistant

As shown in Figure 4, there is a wide range of jobs and skillsets that take part in digital preservation. Of this list, researcher, despite being the second lowest in this figure, had the highest average full time equivalent per role by a significantly large margin (OPF, 2020).

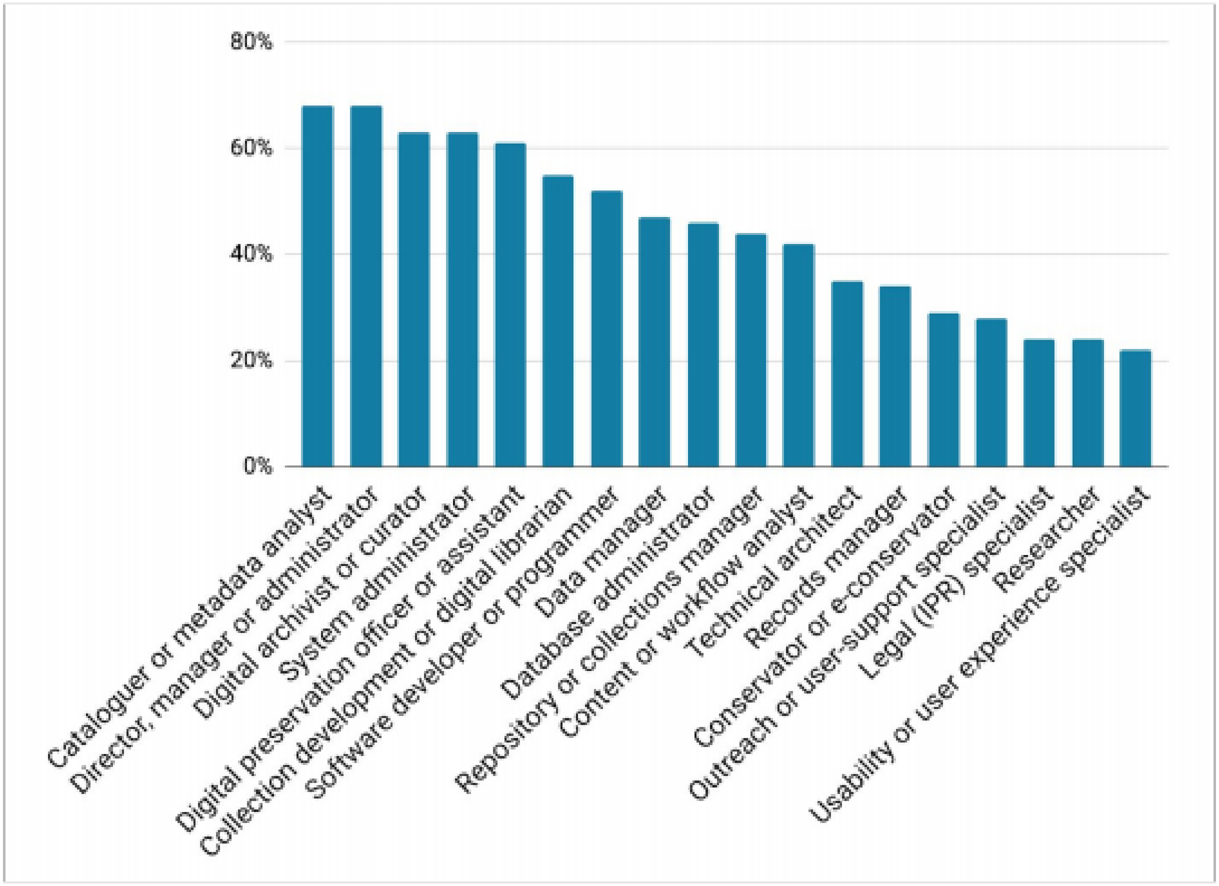


**Figure 4 - Digital Preservation community survey 2019 – 2020 (OPF, 2020)**

There are indications that job roles within collection institutions are both under-utilised and convoluted. The expertise required to improve digital preservation is seemingly present but may not be allocated efficiently. Resource limitations may also be a factor as certain roles are taking on additional tasks outside of their defined title, which divides their focus. The definition of curator roles is diverse across different regions, therefore, the tasks they perform are likely to be affected (Tammaro et al., 2019).

Many of the issues listed above were conceived as key threats in the early study of Jones and Beagrie, (2001) which were reviewed in 2015 from the perspective of Australian collection institutions (Harvey, 2015). At the time, these threats were still valid. From these threats, the following are considered relevant to the issues addressed on awareness:

- Lack of awareness by stakeholders
- Lack of the necessary skill sets
- Lack of agreed international approaches
- Shortage of practice models
- Lack of ongoing funding

- Lack of agreement about who should preserve digital materials
- Lack of applicable selection principles

The list of threats addressed by Harvey and his perspective on the need for collaboration as well as acknowledging the legal aspect of collections aligns with many of the principles in this thesis.

With training and experience, staff performing curation and preservation tasks will have the required information and knowledge to be aware of potential issues and threats. Awareness will enable practitioners to seek solutions and improve the chances of discovering the right tool or method for their needs. Trained staff members with diverse skillsets will benefit from and be benefits to collaborations with other institutions. They will be better suited towards knowing when collaboration is required. Staff that are not performing technical activities may still suggest the best delegation of resources to aid in preservation requirements should they have the knowledge and training to do so. Collection institutions may be required to re-assess their staffs existing skillsets to better utilise their expertise.

## 2.5 Workflows

Workflow diagrams can range from a basic flow of processes that simply show the order in which processes are conducted, to a concise business process model (BPM) that has been derived from the analytics performed with process mining tools (Aguilar-Saven, 2004; Business Process Modelling, 2018; Melão and Pidd, 2000).

BPM  has become one of the main methods for analysing and maintaining business activities, becoming more predominant in influencing decision-making management in small to medium organisations (Grigorova and Mironov, 2018). Grigorova and Mironov state there is no universal standard for creating BPMs, but two of the most prominent standards that exist today are the Business Process Model and Notation (BPMN) (OMG, 2011; Zarour et al., 2019) and the Event-driven Process Chain (EPC) (Software AG, 2020). The authors further express the issues surrounding not having unified standards and suggest the conversion of existing BPM models to workflow patterns that “describe the behaviour of business process” (Grigorova and Mironov, 2018).

The intent to use workflow management systems (WfMS), for the purpose of automating business process execution, can determine how workflows are modelled based on available standards (Ferme et al., 2017). Ferme et al discuss the unfulfilled expectations and the lessons learned regarding WfMS. They recognise the need to develop workflows based on standards such as the Organization for the Advancement of Structured Information Standards (OASIS, 2018) and their Web Services Business Process Execution Language (OASIS, 2007) as well as the BPMN. By developing workflows based on these standards it increases the selection range of WfMS. However, having multiple standards and WfMS can lead to inconsistences, limited support, and pitfalls regarding usability, reliability, and portability (Ferme et al., 2017).

There is something to gain from reviewing digital forensic process models (DFPM). Kohn et al., (2013) reviewed six influential DFPMs from different authors. Each model is presented in the following notation:

DFPM = {start ⇒ next ⇒ then…end}

|| - Indicates parallel processes

^ - Indicates returning to a previous process defined in { }

⇔ - Indicates where a previous process can be repeated after executing the current process (Not used in provided examples, but can be found in the source material)

To show an example, the first model analysed was formulated by Henry Lee and was described as a "Scientific Crime Scene Investigation Model" (Lee et al., 2001). The model was constructed as follows:

**Lee** = {Recognize ⇒ Identify ⇒ Individualize ⇒ Reconstruct}

Where

**Recognize** = {Document ⇒ Collect and Preserve}

**Identify** = {Classify ⇒ Compare}

**Individualize** = {Evaluate ⇒ Interpret}

**Reconstruct** = {Reconstruct ⇒ Report and Present}

This model is focused on physical evidence and has therefore been criticized as it does not consider digital evidence, however, Kohn et al., (2013) state it can be adapted to include digital evidence. The other five models were analysed and discussed in the same way, leading to the Integrated Digital Forensic Process Model (IDFPM). As there were surrounding issues with the existing models and the terminology within, the IDFPM is more than just a merging of these models, it "purifies" and standardises the terminology. Expressed in notation, the IDFPM is described as follows (see Figure 5 for diagram):

**DFPM** = {{Preparation ⇒ Incident ⇒ Incident Response ⇒ Physical Investigation || Digital Forensic Investigation ⇒ Presentation} || Documentation}

Where

**Preparation** = {Policy/Procedure ⇒ Operational Readiness || Infrastructure Readiness}

**Incident** = {Detect ⇒ Assess || Confirm ⇒ Notify ⇒ Authorize ⇒ Deploy}

**IncidentResponse** = {ApproachStrategy ⇒ Search ⇒ {Recover || {Seize

⇒ Preserve} || Preserve} ⇒ {Transport ⇒ Store ⇒ Collect}}

**DFI** = Collect ⇒ Authenticate ⇒ Examine ⇒ Harvest ⇒ Reduce

⇒ Identify ⇒ Classify ⇒ Organize ⇒ Compare ⇒ Hypothesize

⇒ Analyze ⇒ Attribute ⇒ Evaluate ⇒ Interpret ⇒ Reconstruct

⇒ Communicate ⇒ Review ^ {Reconstruct ⇒ Hypothesize}

**Presentation** = {Report/Present ⇒ Decide ⇒ Dissemination}

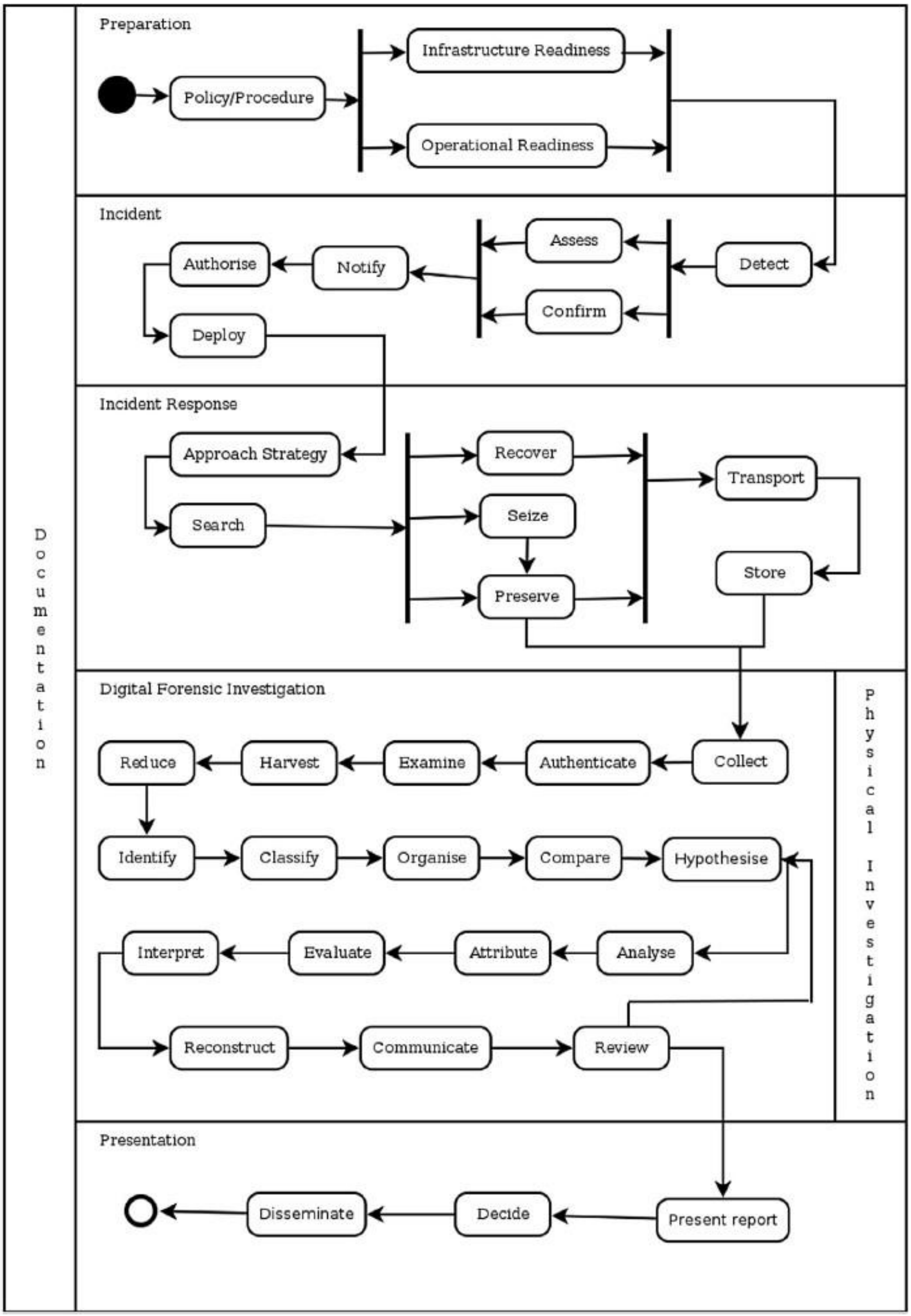


Figure 5 - Integrated Digital Forensic Process Model (Kohn et al., 2013)

The similarities between a digital forensic investigation and digital preservation are undeniable. Both are concerned with careful collection and analysis, ensuring the data are not mishandled and can be presented; one in court, the other in a public collection. Whilst this changes the content that is revealed, as the forensic analyst is concerned with the sensitive data, the collection institution may wish to carefully manage access to such information. The thoroughness in finding information for use in forensic investigations is a practice that will benefit collection institutions. Thus, digital forensics processes and workflows can help in strengthening parts of a preservation workflow that deals with the discovery and handling of sensitive data.

Whilst high-level workflows act as an overall guide to an institution's process, the nature of digital preservation requires adaptability. The discussion in the Bentley Historical Library (Eckard and Hagen, 2018), mentioned in section 2.4 Tools, discusses the approach of mixing custom workflows for each unique collection and a single one-size-fits-all workflow solution. This, on top of adopting an iterative methodology became an essential aspect of their process. This allowed flexibility in how they met their milestones. They incorporated interviews with their archivists as well as those from peer institutions and they emphasised frequent and direct communications among the project team and others involved, leading to new ideas and solutions.

Eckard and Hagen, (2018) stated that whilst researching digital preservation theory and best practice was valuable, it was the "face-to-face" encounters that had the most impact. Best practice should always be a consideration, but institutions should not let it stand in the way of taking action as flexibility, iterative and active processing, and the willingness to adjust workflows are essential (Schroffel et al., 2018). Workflows will remain ad hoc for some institutions, especially small-scale institutions that do not have a consistent volumes of digital material intake (Post et al., 2019).

Solutions may stem from collaborations that may not suit best practice or typical procedure, but these standards do not have the inside information that the staff working at these institutions have.

> *"[...] new standards and best practices are developed in many different areas of bibliographic control, it is impossible to expect one person in the unit to have all the knowledge and expertise. All professionals in the cataloguing and metadata services unit have to work together to stay current on new rules and best practices, and decide*

> *on local policies that would work best for each institution's specific needs."* (Han, 2016)

Therefore, talking to one another can lead to more tailored solutions which should be considered above standard or global best practice. When developing and visualising these workflows; communication, collaboration, and publication should be kept in mind and encouraged (Willoughby and Frey, 2017). This will lead to developing new best practice solutions specific to the institution, rather than following solutions that may be best for some, but not for others. The results of collaborative constructed workflows were revealed amongst the OSSArcFlow partners, allowing them to reflect on their practices with feedback commenting on the insight gained and the utility of the workflow models (Post et al., 2019).

> *"The production and definition of the "archive" must become collaborative in a co-creation enterprise or what has been described as a "curated conversation" that extends well beyond the existing customer base."* (Moss and Gollins, 2017)

Developing partnerships, committing to new and experimental projects, actively revising workflows, and developing policies to support this mediation are supported as benefits in Lampert and Vaughan, (2018). Lampert and Vaughan discuss the activities that took place from 2001 – 2018 at the University of Las Vegas, Nevada, and state the steps involved in their strategic plan only became truly effective with the formation of "interwoven partnerships" comprised of key faculty and staff. Whilst having partnerships and collaboration is essential to improving preservation goals, it may also be a necessity in meeting them, especially in small to medium sized institutions. Management and preservation of digital assets is not easily handled if the resources are not available, as discovered by the Atlanta University Center Robert W. Woodruff Library. This comment is specifically about forming a group to "develop and maintain platforms and systems for preservation and display of online resources". Therefore, the author recommends that "institutions must look toward collaboration both internally and externally to succeed in providing sustained online access to unique digital resources" (Wiseman, 2016).

Collaboration benefits will vary and may have a greater impact depending on the nature of the collection institution. Walters and Skinner, (2011), regarding research libraries, stated the growth rate of "intellectual objects" is increasing at an alarming rate. Collaboration is one of the main recommended solutions, highlighting the necessity of collaboration to meet the large and ever-shifting challenges. Several years have passed since this report was published and the growth rate of data has increased exponentially. The availability of resources within collection institutions has not grown enough to meet these demands, hence the backlog of materials in need of processing likely to be found in all collection institutions.

Furthermore, collaboration is essential in expanding the knowledge and understanding of other disciplines and being aware of what they can offer to the overall improvement of digital preservation. Some tasks may be specific to the preservation aspect, but there are also archiving and curation activities that are involved. Langley, (2020, 2018) argued for a more holistic approach, encouraging collaboration between disciplines with complementary skill sets. With this, the term “Digital Stewardship” was suggested, allowing a broader perspective and collaborative approach to the long-term management of digital content.

Digital preservation needs to flexible and ready to adapt to new technologies and data trends. It therefore needs to be ready to adopt new methodologies and solutions. This is achievable through collaboration with those already partaking in these activities.

## 2.6 Maturity Levels

Collection institutions are at various stages of maturity regarding their digital preservation capabilities. A low-level maturity is not an indication that an institution is inadequate or performing poorly. Many institutions may still be in their infancy regarding born-digital preservation with low demand and low volumes of digital materials. Collection institutions develop policies and implement systems when needed when resource constraints need to be enforced.

Maturity levels are referred to throughout this study, specifically regarding the Australian institutions. These levels are based on public information and the data provided in the questionnaires returned from participating institutions.

Maturity levels are not being used to assess or judge collection institutions, but prompt considerations such as resource restrictions, training requirements, implementation, and overall complexity of workflows and workflow diagrams. The assignment of low or high maturity level is determined by public knowledge and communications with participating institutions and assessing this information against the maturity models. Institutions that have chosen not to participate as they believe they cannot provide any of the required information based on the current state of their preservation levels are considered low maturity.

There are several models for determining maturity levels, however, there are three models that have been selected based on their presentation, influence, and organisational association, to serve as a basis to loosely influence the evaluation of the collection institutions within Australia. The term “loosely” is used here as the maturity levels of institutions are not impactful to this study, they are, however, possible indicators as to why certain systems or policies have not been developed. If there is an average of low to medium maturity levels, considerable thought must go into the proposed solutions and the suggestions on implementation. If collection institutions have yet to establish a dedicated digital preservation strategy, one cannot expect the institution to adopt and implement digital forensic tools and methods.

### 2.6.1 Model 1 - DPCMM

The first model comes from the Council of State Archivists, the Digital Preservation Capability Maturity Model (DPCMM) (Ashley and Misic, 2019; Dollar and Ashley, 2015). This model has many elements; however, the main areas of focus are the five stages of digital preservation capability. The stages include, from top to bottom: Optimal, Advanced, Intermediate, Minimal, and Nominal.

Stage 1: Nominal

This is the lowest stage in which the institution is aware of the specifications of ISO 14721 and other standards. They have either been accepted in principle or are under consideration. There has been no formal adoption or implementation. There is a basic level of understanding regarding digital preservation issues and concerns. The extent of practices mostly consists of ad hoc electronic record management.

Stage 2: Minimal

This stage describes a surrogate preservation repository being available to satisfy some of the ISO specifications. There is some understanding regarding digital preservation issues and strategies, limited to a few individuals.

Stage 3: Intermediate

ISO specifications and standards are embraced. Best practice and schemas establish the foundation of implemented digital preservation capabilities. Successful projects can be repeated, fostering collaboration and shared resources between units and entities managing and maintaining trusted digital repositories.

Stage 4 assumes few electronic records are at risk, and stage 5 assumes no electronic records are at risk. These are not deemed achievable at this point in time. The scope of the DPCMM is made up of fifteen components from two categories: Digital Preservation Infrastructure and Digital Preservation Services. The fifteen components are:

**Digital Preservation Infrastructure**

- Digital Preservation Policy
- Digital Preservation Strategy
- Governance
- Collaboration
- Technical Expertise
- Open Standard Technology Neutral ("OS/TN") Formats
- Designated Community
- Electronic Records Survey

**Digital Preservation Services**

- Ingest
- Archival Storage
- Media/Device Renewal
- Integrity
- Security
- Preservation Metadata
- Access

The DPCMM model further contains metrics and scoring schemas which can be used to calculate and quantify the maturity levels of digital preservation capabilities, however, are not considerations for this study.

### 2.6.2 Model 2 – NDSA LoDS

The second model, Levels of Digital Preservation (LoDS) from the National Digital Stewardship Alliance (NDSA), is based on the 4 levels of digital preservation with focus on specific elements such as: storage, integrity, control, metadata, and content (Kussmann et al., 2020). This model is presented in a colour-coded table that describes each level for the listed elements.

The four levels are described as follows:

- Level 1 – (Know your content)

- Level 2 – (Protect your content)
- Level 3 – (Monitor your content)
- Level 4 – (Sustain your content)

Integrity, Metadata, and Content are the areas that have been selected to evaluate further to influence how maturity is perceived. Storage and Control are omitted as they are based on standard procedures that should be utilised, not just in digital preservation. This includes having multiple copies of files, with recovery and disaster plans in place, and also authorisation and access procedures to restrict, log, and audit read, write, move, and delete functions.

Regarding integrity, Level 1 involves basic operating procedures such as virus scanning and quarantine procedures. Levels 2 and 3 of integrity involve being able to verify and check the integrity of information when handling it. This includes the use of write-blockers and being able to verify files at fixed intervals. Level 4 is reached when fixity can be verified to specific events and actions as well as being able to repair or replace content in the event of integrity loss. Levels 3 and 4 are optimal and indicate a higher level of maturity. If an institution is functioning at the second level, this implies a basic understanding of integrity. This may also imply the institution is growing and may in fact be able to progress to Levels 3 and 4 with the appropriate tools and training.

Metadata is a key element to digital preservation. Institutions without dedicated metadata tools or digital forensics influence are likely to be functioning around Levels 1 and 2 of this model. Level 2 specifies the storing of enough metadata to know what the content is. The metadata elements may be a combination of administrative, technical, descriptive, preservation and structural metadata. Level 3 requires decisions on metadata standards and how any gaps can be filled to meet the standards. Level 4 is met if the institution can record preservation actions associated with content as events occur. The implementation of metadata standards, such as Dublin Core and PREMIS, is part of this level.

Regarding content, Level 4 is only achieved by institutions with dedicated digital preservation workflows. This includes being able to perform migrations, emulation, normalisation, and all related activities that ensure meaningful access to preserved content. It is believed that most collection institutions are performing some of the actions of this level. Level 3 may also be conducted to some extent as it regards the monitoring of obsolescence and changes in technologies on which content is dependent.

### 2.6.3 Model 3 – DPCRAM

The third model investigated is the Digital Preservation Coalition Rapid Assessment Model (DPCRAM), referred to as the "RAM" model. (Mitcham and Wheatley, 2019). This model was designed to be applicable to organisations enabling benchmarks, comparisons, and the ability to contrast their maturity levels. The primary mission, approach, and scale of an organisation do not restrict the use of this model.

The utilisation of the model is not a consideration for this thesis, but the definitions of the maturity levels are. The RAM model is separated into two groups: organisational capabilities and service capabilities. Each capability is evaluated on a scale of 0 to 4. These consist of: Minimal awareness (0), Awareness (1), Basic (2), Managed (3), and Optimised (4).

Based on each of the capabilities and the examples provided for each scale, the following have been selected as considerations for evaluating maturity levels as they directly relate to the areas of improvement targeted for this study:

**Policy and strategy** – The governance of operation and management.

**Legal basis** – The management of contracts, licenses, legal rights, and responsibilities in relation to acquisition, preservation, and access.

**IT capability** – IT support for digital preservation activities.

**Acquisition, transfer, and ingest** – Processes, Donor relationships and communications.

**Bitstream preservation** – Storage and integrity of digital content.

**Content preservation** – Preservation of meaningful and functional accessibility.

**Metadata management** – Creation and maintenance of sufficient metadata to support preservation, management, and usage.

### 2.6.4 Model Usage

The listed capabilities relate to areas in which digital forensic enhancement may improve or depend on. Whilst the goal is not to change existing core preservation-based procedures, the level of maturity for such capabilities does impact successful implementation. If a collection institution has a low maturity level for "bitstream preservation", it suggests there may be complications in implementing digital forensic tools and methods. For example, the RAM model assumes "bitstream preservation" to be at a Level 3 maturity (managed) if content is managed with integrity checking and replication to one or more locations. This level also includes authorisation enforcement for access and suggests tests are routinely conducted to verify effectiveness of backups, replication, and integrity checking.

The acquisition, transfer, and ingest capabilities refer to donor relationships and communication, of which are critical stages for when the prevention of legal and ethical issues begins. Therefore, the overall maturity level of a collection institution is relevant in determining the suggested methods regarding the implementation of digital forensic enhancements.

Maturity models are designed with scoring and calculations in mind. However, for this study a simplified method of evaluation is proposed to loosely gain an estimate of the maturity and performance level of each institution to determine if the suggested enhancements are viable. The models provide a strong influence for an overall consideration as to what to evaluate. For the institutions of Australia, the main concerns are dedicated preservation strategies, software and tools, and the understanding of issues regarding sensitive data.

Based on the selected models, the levels of maturity considered consist of "high" and "low". Low maturity levels align with the lower-level scales of the three models and high levels of maturity will be considered for institutions that align with the highest and second highest levels from the models. Collection institutions with low levels of maturity may be referred to as being in their infancy regarding digital preservation.

## 2.7 Summary

Through research and investigation into existing literature and online resources, the comparisons and differences between digital preservation and digital forensics have been explored. When comparing digital preservation with digital forensics, the criminal and legal nature of forensics is addressed. A criminal investigation has a finite lifespan, a start and end date. Whilst an investigation may span several years, once the case is closed, the data will not undergo any further processing or maintenance. Here lies the main difference between the two fields as digital preservation is concerned with continued and long-term access to the data in which they preserve.

Digital forensic investigations are often conducted on modern devices and media. Digital preservation will always be behind in this regard, often dealing with legacy media and devices. Therefore, the software and hardware developed for digital forensics, whilst beneficial for preservation, are not developed with preservation goals in mind. This makes it difficult for the preservation community to be aware of potential forensic solutions that can benefit their institutions.

Digital forensic solutions open a range of new issues, specifically issues regarding ethics, privacy, and legal. Collection institutions within Australia are exempt from the Australian Privacy Principles, however, with the implementation of digital forensics, sensitive data that were not previously discoverable can introduce new issues that exemptions do not cover. Although exempt, the Australian laws surrounding information privacy and collection institutions can be considered as guidelines, even if they are not enforceable. Identifying where exemptions can be overruled suggests which laws should be taken under advisement.

The potential of digital forensics and the impact sensitive data can have on collection institutions are significant ethical considerations as donors and donations are a primary source of collection data. Donors' privacy and the privacy of their relatives are at risk when sensitive data are discovered. Medical history is such data as it is possible to deduce health conditions on living descendants if a hereditary disease was discovered. Other deductions are possible based on sensitive data discoveries that can infringe the privacy of others.

Provenance is another key aspect to both digital preservation and digital forensics. Digital preservation requires authenticity and accuracy of their collection items. It is also important to understand the thought process in the creation of collection items and how they are meant to be viewed or handled. The history and creation process of an item, as well as information about the author and their mindset are all contributing factors in ensuring accuracy in the display of collection items.

Differentiating originals from copies or fakes is necessary within a collection institution. This differentiation is equally important during a criminal investigation. The provenance of data can establish a chain of custody which is essential when trying to convict a felon based on the items in their possession. If an item of interest is found on the personal device of a suspect, there are various ways in which illicit material can find its way into the possession of unsuspecting users. Therefore, provenance is essential in establishing the relationship between the suspect and the illicit material.

There are many digital forensic methods and techniques that allow digital preservation to be conducted more thoroughly. Ensuring data are not changed or compromised is essential to both digital forensics and digital preservation. Digital forensic tools such as write blockers and checksums (hashes) allow these issues to be discovered and prevented. With the creation of disk images, the integrity of the original data remains intact. Automation and the ability to review and filter information are benefits of digital forensic tools that preservation workflows can benefit from. Sensitive data discovery not only allows ethical, privacy, and legal issues to be addressed, but may also provide new and beneficial information to a collection, not discoverable without the aid of digital forensic methods. Discovery of system environment data for emulation creation and authenticity is also considered essential and the digital forensic tools explored have shown to be useful.

Some collection institutions already have digital forensics solutions as part of their preservation process. Others are either researching or are in the early stages of adoption, and there are also institutions that are not yet considering digital forensics. It is evident that there are many tools and solutions, however, there lies a problem in inconsistency. Given the nature of born-digital data which have many unique properties and are quite volatile, this is to be expected. One of the key reasons behind this inconsistency is awareness. Awareness of solutions to existing issues, as well as awareness of the issues themselves. The solutions used on existing issues may cause others to emerge. Without being mindful of potential outcomes, the likelihood of something going wrong is increased.

Other factors such as differences between collections, their priorities, mandates, and other forms of restrictions contribute to the inconsistencies. Awareness, however, has the potential to make a significant difference. For example, if an institution needs to perform a task and is only aware of a single solution, this may be problematic if the solution exceeds their budget or staff resources. The solution may also require training of dedicated staff, which too will incur cost. If there were a solution that meets their requirements without exceeding their budget, being aware of such a solution would greatly benefit the institution.

Finding an appropriate solution amongst the many digital forensic tools and methods can be determined by their accessibility and discoverability. Documentation and how these tools are presented are factors in their discoverability. This highlights the disregard for any potential functionality outside of a tools primary purpose. Whilst attempts are being made to close the gap between digital forensics and digital preservation, if digital forensic solutions are not presented in a manner desirable for preservation, collection institutions remain unaware of the potential and the necessity for digital forensics implementation.

One of the contributing factors regarding awareness issues is the quality of digital preservation workflows, specifically their design and transparency in the processes that are visualised or omitted. Workflows themselves are not treated equally across disciplines. Some organisations use workflows as a dedicated business process model that is strictly followed. Others, such as digital preservation, require workflows to be flexible and amendable as each preservation case may have unique properties and requirements. Therefore, preservation workflows should be considered as guidelines. As digital preservation requires human intervention in most stages of processing, workflows are not treated as an automation scheme.

There are many variants of workflow notation and design. The structure of workflows can be quite different based on the creators modelling experience. The layout and presentation of workflows are often unique per institution or group. Complexity is also a considerable variable. One must realise that not all people performing digital preservation may have a technical background and may not be familiar with Unified Modelling Language (UML) used for the modelling and notation of workflow diagrams. Modern workflows are often created linear, reading left to right, making them easier to follow. Earlier models, however, were designed top-down, resulting in higher density of node clusters, which at first glance can seem complex. These differences were present between the 2012 and 2016 U.S datasets used in Sections 5.1 U.S Collection Institutions – Tools and 7.2 U.S Collection Institutions - Workflows.

Many librarians have migrated into a preservation role based on the needs of their institution. Therefore, workflow complexity is an issue that must be addressed as it should be interpretable by all levels of experience.

As transparency is considered a requirement for peer-to-peer collaborative learning and improvement, the institutions that have been assessed have been assigned a maturity level. These maturity levels provide a baseline for each institution and how developed they are in their preservation operations. Many factors are considered when determining maturity levels. Three models have been described with varying criteria by which maturity is determined. Maturity levels are used as an estimate between early adoption of digital preservation and having a dedicated preservation workflow with digital forensics implementation or consideration. The differences in maturity levels may differ based on several key variables. Intake, resources, demand, policies, and other limitations may factor into maturity levels. For example, one institution may only accept government material, whereas another may accept all material related to their state of origin. The size of an institution and their funding allocation are also significant factors.

Differences in maturity levels are expected and will continue to exist as each organisation progresses at its own pace. The issue is when institutions with a lower maturity level are looking to their higher-level peers for guidance. When an institution at a high-level of maturity is effectively utilising dedicated tools for preservation and has successfully implemented digital forensics, and if they are transparent and accurately visualise their workflows, these resources can serve as learning tools and guidance for others.

Transparency is not only important for other institutions, but for donors as well. A donor should be able to access an institution’s website and see their digital preservation policies and workflows. This will help them decide as to whether an institution is suitable for their material if they possess the knowledge to interpret such information. Donors should know exactly how their material will be handled as well as changes in ownership and rights.

The differences and similarities between digital preservation and digital forensics have been addressed and considered throughout this study. Sensitive data discovery forms the main objective for the investigation into possible solutions. Ethical, privacy, and legal issues are constant factors when evaluating the potential of digital forensic tools. Lack of awareness and transparency are obstacles to overcome for many institutions and is one the primary motivators behind the solutions discussed.

# 3 METHODOLOGY

This chapter explores how each part of the methodology was derived. This includes the actions that were taken, the thought process behind the methods, and the results that led to new methods being established. As each discovery was made, new ideas formed, and this chapter is structured to describe the process. Each sub-heading reflects a step in the methodology.

## 3.1 Gathering Public Data

In order to establish what information was publicly available with respect to digital preservation policies and procedures an initial investigation was undertaken searching online documentation of Australian collection institutions. However, given the large number of collection institutions within Australia, the institutions with the most publicity and influence were selected. The state or national libraries for each state and territory were selected, resulting in eleven institutions. The collections within these libraries hold data of great significance to their respective state or territory, including government and national historic data. One would see these institutions as the exemplars and would assume smaller libraries would look to them for guidance on policies and future endeavours. The ideology is that these institutions should be at the top of their field with more resources than smaller, private institutions, and should therefore be setting the right example.

This initial investigation involved searching the websites and online documentation for each selected Australian collection institution. Through this method, the available information was discovered, establishing how effective it was at conveying the desired information, and what was missing. The following criteria were used:

**C1** - How easy is it to find the required information? How many mouse clicks/breadcrumbs are required? Are the policies accessible from the home page?

C1 is based on how easy it is to find information regarding digital preservation policies and donor agreements. This includes how many mouse clicks are required and whether this information can be accessed from the home page, e.g., in one mouse click.

**C2** - Are all the policies stored in a single location?

C2 determines if all the information can be found in one location, e.g., one page with links to all the policies and supporting documents.

**C3** - Do the policies include: digital preservation, donor agreements, and any supporting documents?

C3 identifies if the required information is among the policies, e.g., are there digital preservation policies, donor agreements, and any relative supporting documentation.

**C4** - Is the digital preservation policy unique to the institution?

C4 is determined if the preservation policies are unique to the institution or borrowed from elsewhere such as a generic list of standards.

**C5** - How informative is the policy? Can adequate information be derived about the institutions position with digital preservation as well as their process/workflow? (includes software/hardware usage)

C5 is based on how informative the policies are, do they give enough information?

**C6 -** Can donors establish how much control they have over their material from the donor agreement documentation, without communication with the institution?

C6 is determined by the donor agreement, how informative and detailed it is, and whether donors can establish how much control they have over their material without having to contact the institution.

It was apparent early in the investigation that much of the information sought after was not included in the online documentation. Information such as the specifics of the digital preservation workflows, the tools and methods used, and any digital forensic techniques were lacking. The information that was present was barely nominal and often directed to a generic external source. The gathered information served as both a comparison and evaluation of the various collection institutions presented by comparing each library and how much transparency was in their documentation. Any information regarding digital preservation policies and procedures was the primary focus. The donor agreement forms, and procedures were also of great importance, determining the level of stipulations and control the donor may retain. Specifics on hardware and software used, as well as details on any digital forensic influences were also sought after.

This provided insight into the potential maturity levels of each institution. This led to a new approach being adopted to gather data, involving direct communication with each willing institution.

Investigating institutions outside of Australia was more rewarding. Whilst there was not an abundant amount of information available, it was possible to find information regarding digital preservation workflows. This information was derived from the BitCurator Consortium repository containing workflow diagrams for several universities and archives from the U.S that represented each institution's preservation process (BitCuractor Consortium, 2018). The accompanying literature helped fill in the gaps, which is where some of the workflows were published, derived from interviews with the respective institutions.

This study acknowledges the differences between the institutions targeted for data analysis and the institutions targeted for enhancement. The mandates, resources, and legal context are different for U.S university libraries and archives from those of the Australian national and state libraries.

The selection of these institutions was based on factors that align with the goals of this thesis. The data had to be publicly and freely available and with a strong influence of digital forensics, as is required by the Australian institutions. The workflows presented by these institutions required the use and acknowledgement of digital forensic tools and methods for sensitive data discovery. The BitCurator Consortium included other partners and institutions, but at the time, these were not open to the public. Furthermore, not all institutions within this dataset are equal and maintain different collections. The institutions from these datasets are varied in size and what they specialise in, providing a range of workflows. This was ideal given the range of data types the state and national libraries receive and preserve.

Over time, additional sources were published, resulting in a collection of twenty-four different workflows, diverse in both design and levels of detail, making up an effective dataset for analysis. At this point, there were four datasets which included the Australian dataset, the U.S (2012) and U.S (2016) datasets, and the OSSArcFlow (2018-2019) dataset. The U.S datasets were combined for the tool analysis and the OSSArcFlow workflows were combined the U.S workflows for analysis.

The comparison of these different datasets was relevant due to four main reasons. Both sets of institutions perform digital preservation, the foundation of this study. The U.S institutions targeted are part of communities that use and acknowledge digital forensics and sensitive data. This was evident by the affiliation with BitCurator and the digital forensic tools and processes within their workflow diagrams, making them prime candidates for goals this study aimed to achieve. The workflows for these institutions were publicly available and more informative regarding their preservation procedures. The Australian institutions that have been reviewed are lacking in this area by not having workflow diagrams or efficient detailed reports on their processes and methods publicly available.

A large portion of the U.S workflows discovered online were often high-level and lacking in the details sought after. This includes how sensitive data are handled, if discovered, and if this is still a consideration after the ingest process once data are stored. Therefore, if these criteria are to be met, there needs to be evidence of:

- Tools or methods that enable sensitive data discovery
- Workflow processes indicating sensitive data discovery
- Processes indicating what happens once sensitive data are discovered
- Checks and stops that prevent workflows from proceeding until these processes are performed
- Considerations regarding sensitive data that may reside in stored content.

There were a small number of exemplary workflows, that to some extent, met these criteria.

Whilst reviewing the workflow diagrams design and effectiveness for conveying the information desired, ideas of additions and improvements that could be made started to develop. As the reviews continued, it became apparent that the requirements were not met in the discoverable material. There were however some institutions that had adopted digital forensic methods to discover sensitive data, with some showing minor details of how these data were handled.

It is worth noting, that any workflows that contained the process of searching for sensitive data, the next step in the workflow often proceeded regardless of the outcome. There was little to no decision-making, terminations, or alternative processes in many of the diagrams visualising how these data are handled.

Collection institutions would take the data they want to make public and give it public access, but what happens to the sensitive data not deemed suitable for public access was rarely reflected in their workflows and documentation. Questions regarding if sensitive data are lying dormant in collections and if those data still have a purpose yet to serve have driven the data analysis and questionnaire development. By analysing the capability of each institution in being able to identify sensitive data effectively, by determining if they have the required tools and methods in place, reveals the potential of stored data that remains unidentified. This helped form the questions that make up the questionnaire in order to ascertain the information required to make this determination.

In fact, some of the responses from the Australian institutions stressed that no redaction or deletion takes place, unless necessary. This view on redaction and deletion gives merit to these questions as the likelihood of sensitive data existing is increased.

Furthermore, emphasis is placed on raising awareness of the existence and importance of sensitive data and the issues that surround it, allowing the prevention of future issues that could eventuate from handling such data incorrectly. Therefore, it is necessary to identify where data are handled incorrectly, such as data that passes through a workflow without the appropriate processing, which is then stored within a collection. The analysis of policies, procedures, workflows, and other related information provides insight into this issue. If the tools or methods are not available, and workflows visualise no error checking or handling, then one cannot assume data are handled correctly.

Prior to the questionnaire development, an analysis of public data was conducted by searching through all the available documentation to establish what tools and methods were being used in the evaluated digital preservation workflows.

## 3.2 Analysis of Public Data

Two sets of workflows were gathered from the BitCurator Consortium, one from 2012 and the other from 2016. Each set was evaluated based on their design complexity, consistencies across each of the workflows within the set, and their overall flow. For example, the 2012 set was less consistent in design and had less of a linear flow. The 2016 set was presented in a design that is much easier to read and follow by making use of swimlanes to simulate which user or system was handling each process within the workflow (Lucidchart, 2019). These sets, from different years, illustrate growth in design and process across similar institutions, revealing the influence of digital forensics on digital preservation from 2012 to 2016. There were no overlaps in institutions across the two datasets, although a single university submitted workflows to each dataset, however, they were from different departments.

The analysis was conducted by following each node, the visual representation of a process or step in the workflow and taking note of every tool listed within those nodes. With this method, a list of tools was generated and put into a spreadsheet. Once all the information was extracted from each workflow and their design evaluated, a combined list of all the tools and the number of times they were used throughout the set of workflows was calculated. See

Appendix C – Tool Data. This enabled a visual representation via tables and bar charts, showing the usage of each tool and the frequency of its use across all the workflows within the dataset. Visualising the datasets separately allowed comparisons to be made on the different tools used and the frequency of their use. Combining the datasets allowed tools to be counted across both sets, revealing which tools had an increase or decrease in usage across the four-year timespan. Applying filters allowed tools with single uses to be visualised as well as those that were used in both the 2012 and 2016 datasets. This provided visual perspectives that allowed new data to emerge.

Focusing on the tools used provided valuable data that could indicate the types of processes and operations involved with the digital preservation workflows. Instances where a tool is listed without a supporting process, as well as a process being listed without the appropriate supporting tool, may give indication as to whether sensitive data are discovered and handled adequately. However, one must be cautious as the existence of a tool may not be an indication that the respective process is in fact conducted. This theory was confirmed with the Australian institutions as they had listed tools that were not made use of, presenting inconsistencies between the tools and processes listed.

It is worth re-iterating the acknowledgement that these workflows may not be an accurate representation and some information may be omitted. If a workflow does not reflect a particular tool or process, it does not necessarily indicate the institution is not engaged in this activity. However, as this presents another key issue that is addressed within this study, “transparency”, they are treated as accurate representations. It does, however, not make sense for information to be omitted surrounding sensitive data given the platform they were published on and the reason for doing so was influenced by digital forensics.

The purpose of gathering and analysing the public data was to establish a comparison with the Australian data collected. This comparison was primarily used to show which tools were shared across the two U.S datasets and the Australian dataset and those that were not. By analysing these data, it revealed which tools had high usages across the three datasets, spanning over several years, indicating the potential benefit of such tools. It also raised the question as to why certain tools, considered to be effective and widely used, would have such a low usage count across the dataset, warranting further analysis. These instances may be due to the institution’s unique position, they may in fact deal with certain media that other institutions do not and therefore require unique tools. This information was determined by correlating the tool data with the data provided on the types of media within the collections of each institution, established via the questionnaire.

The tool data further provided variables that were not obvious to begin with. Variables such as which tools were phased out, leading to questions such as why were they phased out and what replaced them? Researching the tools that had replaced those being used in earlier workflows revealed some indication as to why this transition occurred. This was clear in some cases as the replacement tools had superseded their predecessors. Through this method, discoveries were made revealing some institutions were using outdated tools.

Whilst the tools an institution uses is something that can change at any moment, there are consistencies such as such as the ongoing use of legacy tools that have been superseded. This may be more prominent in some institutions more so than others, which may be a result of several variables. The common variable in this case being resource and funding limitations.

Whilst the tools may change, the core processes of preservation are not likely to undergo any significant change as these are common requirements that can be achieved by several different tools and methods. For example, disk imaging is common process within a preservation workflow as it prevents the original source from being used. This mitigates any damage or further degradation. The tool to perform this task may change frequently, but the core process itself does not. Surrounding tasks may change, such as including write blocking or checksums, but a disk image is still being created.

Therefore, based on this ideology, the data evaluated are treated with relevancy even if changes occur after evaluation as the patterns and processes are still relevant, even if how they are conducted has changed.

From the evaluations conducted at this point in the methodology, a better understanding of what data needed to be gathered had been formed. The tools, processes, and workflows seen from an international perspective allowed for the development of the questionnaire as discussed in the following section.

## 3.3 Questionnaire Development

The purpose of the Appendix B - Questionnaire was to fill in the gaps where public information was lacking and to gather data based on the findings of the U.S institutions. A documented questionnaire was the selected as the primary approach to gather data directly as it provided the most flexibility and freedom in response, more so than online surveys which are less targeted and more easily ignored or sent to spam. Where possible, known members of the digital preservation team within each institution were targeted directly, based on the information gathered from the correspondence with other members of the NSLA. This turned out to be a strong choice as the data that were returned made use of tables, images, and appendices. Where binary answers where not applicable, the participant had the freedom to express their answer in any way they desired.

At first, the questionnaire gave the participant the option to anonymise their institution which became a popular choice. It was then decided to anonymise all the results using unique identifiers and withholding the name of certain tools and services as they could be linked back to their respective institutions.

The questions were broken up into three parts:

- Donor agreements and legal and ethical standards
- Digital preservation
- Digital forensics

The first section covered the donor agreement process, aiming to identify how ownership, access, and donor stipulations are determined. This section also covered the processes involved when discovering sensitive data that had been addressed by the donor agreement and for data that were not.

The final question for this section asks about the procedures in place should the donor and next of kin no longer be available in the event of sensitive data discovery. It also asks if the type of data changes protocol, for example, if the data were political based or relating to a deceased person. The point of this question was to establish the ethical views on how to deal with data that may be detrimental to a person and or their family's reputation. Being a grey area with collections often being exempt from legal obligations, ethics is a stronger consideration in determining how to handle such a case.

The second section of the questionnaire aimed to establish:

- Data types
- Metadata handling
- Processes and tools
- Workflows.

By establishing the data types and comparing these data with the tools and processes used within the workflows described by each institution, allowed any inconsistencies to be discovered. For example, if a tool is listed with no mention of a supporting process, this is an inconsistency. It also provides supporting information as to why the tools listed are used by examining the data types within the institutions collection. However, it should be noted that many of these institutions are growing and adopting new tools which may not have been implemented yet or the required training has not been conducted.

The third section of the questionnaire aims to determine the understanding of digital forensics each institution had, including if they were already utilising digital forensic tools and methods. These questions were based on what the institution already had in place as well as the options for what could be used.

For example, if the documentation accompanying digital forensic tools were to consider a digital preservation perspective, would this make it more likely to be adopted? Budgetary concerns were also addressed, asking whether open-source, freely available tools were an option where proprietary software exceeded budgetary limitations. Lastly, the final question asked if the institutions were willing to adopt suggestions and improvements to their workflows.

The questionnaire was submitted for ethics approval and was returned with a conditional approval response.  Some changes, mainly clarifications, had to be made and after the first revision, it was approved under the project number **7755**. As required, a letter of introduction and an information sheet accompanied the questionnaire when delivered to the participants. The participants had full access to all the information and the questions before making their decision on participation.

## 3.4 Data Collection

The data collection that followed was focused on the results from the questionnaire sent out to the Australian collection institutions. This was sent out to the nine state libraries and the national library. Note the NSLA was investigated for their publicly available documentation regarding polices, processes, and procedures and any information regarding digital preservation. These institutions were selected as they represent their state and country.

The same process was followed for the state archives, but this did not meet acceptable goals as the responses were not fit for analysis due to being blank, too basic, or refusal of participation due to development progress. The diversity in the results from the libraries alone gave enough of an overall picture of the maturity levels of Australian collection institutions regarding digital preservation. See Section 2.6 Maturity Levels.

Each questionnaire returned was reviewed and compared against the responses from every institution by extracting the information from each answered question as well as any supporting documentation and correspondence provided by the institution. Each answer was summarised, and the key points were listed for each question under the ID provided for each participant in a master document. Through this method, the data were then added to the spreadsheet of existing data from the U.S datasets. This allowed new charts and tables to be created and compared against the existing charts and tables. Combinations of both the U.S and Australia datasets formed new charts to be visualised. This revealed new tools and tool counts.

It should be noted that all data processed went through anonymisation by removing any identifying factors with only the original returned questionnaires containing the unchanged data. The originals are stored and protected in secure online storage.

There were variations in the types of responses received. Some institutions agreed to participate but were unable to continue. One institution refused and another responded stating they were not developed enough to adequately answer the questionnaire. Three institutions stopped responding, despite multiple attempts to re-establish communications via online forms and direct email, and one institution did not respond at all.

Similar results occurred in a survey conducted by DeRidder and Helms, (2016) where a drop in responses occurred. There were 62 participants in the beginning of the survey which then dropped to 20. All 62 answered the first question on what types of data they collect, but responses soon ceased. There were fluctuations for each question, but they maintained an average of 20 responses. It is stated that this is likely due to policies and procedures not being developed past the initial intake.

With the answers provided, the questions that could not be answered, and some of the negative responses from the Australian participants, there was enough data to begin comparing and evaluating where these institutions were in terms of being able to utilise and adopt digital forensic tools and methods, as well as how much digital forensic influence was present. The need for such enhancements was also established. In fact, both the positives and negatives from this process strengthen one or more arguments presented in this study.

## 3.5 Results

After careful analysis of the Australian dataset, derived from the questionnaire and correspondence with each institution, the types of data, the amount of data, the tools and processes used, and the existence or need for digital forensics were established with many variables discovered that could be applied in comparisons with the other datasets. Many concerning discoveries were made, and a better understanding was formed of the different levels of digital preservation across Australian libraries. The following is a list of key discoveries:

- The U.S dataset from the BitCurator Consortium is influenced by digital forensics
- The Australian dataset shows considerations for digital forensics, but is not utilised
- Some tools were shared across the datasets
- Some tools were shared across institutions
- There were many single-use tools
- There are outdated and superseded tools still in use
- The file-types collected by each institution (quantities)
- Basic, conceptual workflows
- The demand and necessity for preservation per institution

This list was established based on whether or not digital forensics is being utilised as expected by focusing on the tools used, which may be impacted by the types of data being handled as well as the demand for preservation, determined by intake. The use of outdated and superseded tools may suggest resource restrictions and budget concerns, all which factor into the derived solutions, considering such variables and limitations.

Inferences made with the Australian workflows were derived from the workflow descriptions provided and by also correlating the questionnaire responses and all the public data collected. The information from each question provided an overall picture of the progression and maturity levels of each institution.

Not only was it possible to identify where improvements could be made in the overall digital preservation workflow, but the questionnaire responses also revealed where some of the institutions wanted and needed improvements to be made.

## 3.6 Digital Forensic Tool Investigation

An assumption was made initially that sensitive data were an issue that was not being approached appropriately. The risks surrounding sensitive data have been established in the literature reviewed and the legal information investigated in Chapter 4 AUSTRALIAN LAW IMPLICATIONS. The data from the questionnaire results and the lists of tools used within collection institutions, found in Chapter 5 WORKFLOW TOOLS – DATA GATHERING, highlight this and strengthen this argument.

Many assumptions could be made about the severity and impact of sensitive data and the ability to capture it. However, when tested by using digital forensic software on real data, that is taking a donated hard drive and processing it with digital forensic software, the results were unexpected, exceeding many of the assumptions made.

The results are explored further in Chapter 6 DIGITAL FORENSICS – SENSITIVE DATA.

The primary goal of the digital forensic tool investigation was to identify sensitive information on a real data source to replicate the risk in both solicited and unsolicited donations. A disk image was created with FTK Imager of a small 13GB hard drive and was used in the experiments. FTK imager was used based on personal preference and being a freely available standalone tool from FTK, a proprietary digital forensics suite of software. Free and standalone are desirable factors as they consider resource limitations as well as training and system requirements that may be needed for packaged suites such as FTK or BitCurator.

The hard drive chosen was selected after a random collection of donations were imaged. No information accompanied the hard drives, only that they were donated to the Digital Archaeology Lab at Flinders University, source unknown. The images were reviewed, looking for signs of active usage with multiple users such as family members or guest profiles. Multiple users generate more data as well as diversity in data, based on how each user interacted with the system. The investigation sought to find how many files, directories, and varying data could be discovered with basic navigation methods such as searching through each directory, using the search function, and looking at the properties of each directory. With this, even a small data source, such as the 13GB hard drive, in comparison to today’s standard 2TB hard drives, can return ample amounts of meaningful data due to how it was used and the frequency of its use. The hard drive used in the experiments was deemed adequate based on the discovery of multiple users and the number of directories and files, compared to the other hard drives.

The first tool used was Bulk_extractor (Garfinkel, 2013). Bulk_extractor was selected based on its free availability, its usage within the workflows evaluated, and personal experience. This software generates an output of text files containing sensitive data such as email addresses, visited websites, search history, and thumbnail images carved from fragmented files to name a few. The output could also be read within Bulk_extractor, presenting the data within the graphical user interface, offering a tailored format. See Section 6.1.1 Bulk_extractor and Regular Expression.

Without discussing the results in detail, the data extracted from the 13GB disk image resulted in approximately nine million (**9,000,000**) lines of data, primarily text-based, with binary and hexadecimal data included, residing in a directory of assorted text files and image files (jpegs). The output of these text files ranged in line counts and the length of each line, with some lines spanning a significant portion of the page, made up of strings and characters which encompassed the key findings. The line count only considers each individual line vertically presented. Bulk_extractor creates histograms, which take only unique values and presents them in a separate text file. For example, it will take all the unique domains and list them, rather than having the same domain repeated. The data are extensive, intrusive, and can reveal a lot about the user(s) as well as the system of which the data source belonged.

The second experiment was a more complete digital forensics process. This involved taking the physical hard drive, connecting to system via write blocking capable SATA to USB device, making an image, both with and without checksums for experimental purposes, and finally processing the image through Autopsy (Basis Technology, 2018). This tool was selected as it represents a complete suite and is a freely available competitor to other popular proprietary software such as FTK and EnCase. Autopsy is the user interface for The Sleuth Kit (TSK) (Basis Technology, 2018a). Autopsy has a much more involved user interface which includes visualisations, graphical representations, and many sorting and filtering options. This involves much more interaction and analysis from the user, but the output is presented in a way that is more meaningful to the user. This requires navigation through the user interface, expanding each directory, and analysing the output displayed when files are selected, as shown in Section 6.1.2 Autopsy (The Sleuth Kit). This includes navigating the gallery where pictures and video files are categorised as well as the visualisation modules that display a timeline of file creation and modification and communications between users.

A third investigation was conducted once the processing of both tools had been completed. This involved crawling through the output data and searching for specific strings of text. To achieve this, a regular expression (regex) tool was used, allowing all the output data to be searched at once with both string matching and regex. This was performed via the external software, grepWin (Küng, 2018) and also with the built in functionality provided by the digital forensic tools used. Known search strings involving sensitive phrases were used as were variations of email regex. Regular expressions are discussed in Section 6.1.1 Bulk_extractor and Regular Expression.

The objective of these experiments was to not only show the potential of the digital forensic tools explored, but to reveal the significant amount of data potentially undiscovered, which could contain beneficial or damaging information to a collection. One cannot assume donors are fully aware of the data that may reside in the media of their donation, nor the digital footprint left behind from their usage of said media. Despite the efforts made during the donor negotiations, precautions must be in place in the event the donor was unaware of this risk.

## 3.7 Data Analysis

The data collected in the digital forensic investigations were analysed using a combination of the third investigation and made use of the user interface supplied with each tool. Manual searching was used to see how tedious and time consuming it could be compared to using search functionality. After several minutes, this was established to be inefficient and ineffective, being slow and unable to locate obscure data or have it represented in an interpretable manner. Manual searching is still required on the output as the data returned can be quite extensive and mixed with false positives, making human analysis an essential part of the process.

The number of elements found for each category was noted and then multiple searches were conducted based on assumed sensitive data targets. This included analysing the output for search history, such as the keywords and or search strings typed into a search engine. The output of keywords displayed how a search engine interprets the input. For example, a search string containing “how to install a cpu” would be read as “how+to+install+a+cpu”. This is how the output from Bulk_extractor was displayed.

Select strings were chosen as search parameters, including known domains for piracy and other sensitive websites. All searches returned no evidence of illegitimate activity; the hard drive was not a part of any illicit online activity such as piracy. Some searches revealed sensitive data such as explicit search history, private correspondence, and personal private photos, but nothing too severe was discovered.

The analysis of the Autopsy data was able to take the search parameters and search within documents that were present on the hard disk drive. This includes text-based documents and emails, providing transcripts of correspondence between the user and associates, as an example. With each experiment, the output was reviewed whilst maintaining a digital preservation perspective. This means there was an additional focus on data that could aid in the creation of a user’s profile, giving context, as well as data that describes the computing environment of the source for emulation purposes. The motive behind this perspective is the ability to provide meaningful access and not just the ability to preserve data. This perspective does share similarities with that of a forensic analyst; therefore, the criminology aspect was not ignored and provided an approach with to test the software.

Conclusively, these experiments provided examples on the data gathering capabilities for both sensitive and non-sensitive data, the ease and speed with which this can be accomplished, and the manner in which information is displayed and visualised. The range of benefits are evident when using digital forensic tools, even if one simply wanted their data displayed and categorised within a user interface that allows them to explore the data more thoroughly. This addresses the first and second research questions regarding greater data gathering capabilities and being able to discover said data in obscure locations.

## 3.8 Workflow Enhancement

The proposed solutions to the issues addressed throughout this study are preventive measures that will provide additional benefits other than threat mitigation and removal. The proposed workflow enhancements aim to enhance, amend, and improve existing workflows, not to completely redesign them. There are too many factors that make it inconsiderate to expect institutions to completely change their core processes, such as resource constraints which can include budget, staff, and training. Due to this, the recommendations were tailored to be modular so that they can be added to existing workflows and implemented at the institution’s discretion.

The method to reach this solution was derived from the investigations and analysis that were conducted. This revealed the depth of data that can be discovered through using digital forensic tools, revealing the range of data that can be harmful or useful to a collection. More information, regardless of how little, can change the nature of a collection item and this may be a positive or negative change.

The data derived from the OSSArcFlow (2018-2019) dataset further strengthened the arguments presented regarding sensitive data and the impact it can and will have on collection institutions. These workflows are evaluated in section 7.2 U.S Collection Institutions - Workflows. With the BitCurator Consortium and the OSSArcFlow datasets combined, the data from the workflows compared against the questionnaire responses provided a baseline for comparison with Australian institutions. The combined analysis used the following criteria:

- Design (use of swimlanes, UML notation)
- Decision-making (appropriate placement and termination when necessary)
- The inclusion of donor agreements (donor interview, accession, etc.)
- The inclusion of sensitive data discovery
- The inclusion of sensitive data handling

Each of the 24 workflows was documented on if and how the criteria were met. The results were tallied and visualised in charts that indicate how many workflows met the criteria.

The handling of sensitive data is of the utmost concern and there are pertinent reasons as to why this is. Consider the following hypothetical scenario:

Institution **A** collects dataset **B**. Dataset **B** is a collection from an author who specialises in poetry. Dataset **B** is made up data **C**, **D**, and **E**. **D** contains a list of poems institution **A** wishes to categorise and display in their public collection. This is done by only making **D** publicly available whilst **C** and **E** sit in storage or are discarded. **C** and **E** could contain sensitive data that can either change the nature of **D**, revealing new information, or it could lead to a new investigation altogether, such as new persons or items of interest, and in extreme cases, information of a criminal nature may surface.

To further the example, **E** could contain an email trail between the author and another individual. The contents of those emails could reveal that the author is not actually the original creator or there may have been a co-author who is not getting credit. The email trail could reveal several key bits of information that change the perspective of the collection items. This threatens the accuracy of the collection as incorrect or incomplete information is published. This creates an obligation to further investigate these claims and to verify their accuracy. If **E** contained traces of illegal online activity, law enforcement should be contacted immediately.

Depending on the nature of that sensitive data, different threats may emerge. Whilst the public may not have access to **C** and **E**, staff members do. Human error or malicious intent are events that can take the threats sitting in storage (**C** and **E**) and exploit or unleash them into the collection. From an optimistic perspective, **C** and **E** could potentially strengthen **D**, making it a greater asset to the institution’s collection. There is either a threat or missed opportunity should the contents of **C** and **E** remain unknown.

The issues surrounding sensitive data are explored further in Chapter 4 AUSTRALIAN LAW IMPLICATIONS.

Once it was established how each institution was performing, the exemplary examples were selected to form a baseline of the improvements to be designed for Australian institutions. It should be noted, that whilst a small selection of workflows strongly met the criteria listed above, there was still room for improvements.

From the evaluations providing examples that may or may not be suitable, new workflows were created based on different parts of a digital preservation workflow. Each enhancement was designed to meet the criteria specified and to ensure a complete end-to-end process for data to safely pass through. This ensures no data are overlooked and integrated into the collection where they can cause harm should they not have been investigated carefully.

Multiple workflow diagrams were developed as some formed sub-diagrams of higher-level processes in order to reduce overall size and complexity of the main workflow. For example, in the main workflow presented, the majority of the data passes through the evaluation node. This node is quite expansive and has many processes involved. To include all this in a single diagram would severely increase complexity. Therefore, this node is presented in its own diagram to allow for freedom in how it is presented.

This is determined to be the best approach for several reasons, the main one being a user-friendly approach to allow users of all levels the ability to read, change, and add to these workflows. As every institution is unique, adjustments will need to be made by the staff and the easier it is for this to be achieved, the better the outcome will likely be. Hence the design of the workflows is generic, free of specific design and notation, to allow better adoption into existing workflow designs.

The key point in understanding the need for such enhancements is to understand the threats of the past, present, and future. The past can reveal the potential threats of the present by determining if sensitive data discovery and handling was performed. If this did not occur, then there may already be dangerous data sitting in storage. If these issues are ignored, this opens potential threats in the near future. Even under the exemption of privacy laws for collection institutions, laws can be re-written and, in some cases, still be enforced retrospectively. Defamation law is an example where exemptions can be overruled as discussed in Section 4.3.2 Defamation.

One could argue the impact of information, regardless of its provenance date, can cause a severe impact in the present. For real world examples, one must look no further than social media. In 2018 alone, many people had their lives and jobs impacted due to evidence retrieved from social media posts published several years ago. These posts may include something being said in poor taste or political views that are now frowned upon. The point is, 10 years ago, times were different, but regardless of how different things use to be, the past is always seen through the scope of today. For example:

> *"Disney has fired writer-director James Gunn from 'Guardians of the Galaxy Vol. 3' after a right-wing media personality resurfaced a series of offensive tweets Gunn made, in many cases from 2009 and 2010."The offensive attitudes and statements discovered on James' Twitter feed are indefensible and inconsistent with our studio's values," Walt Disney Studios chairman Alan Horn said in a statement to The Hollywood Reporter, "and we have severed our business relationship with him.""* (Bishop, 2018)

As can be seen, regardless of the current views Gunn has and how apologetic he is, he still suffered the consequences of his actions in the past. Disney was quick to act, one may argue, prematurely, as the risk of aligning with someone like this introduces the threat of backlash and bad publicity.[1]

Now whilst collection institutions do not have the same exposure and publicity as a famous movie director, and whilst the institution is tasked with providing information which can often be distasteful, as many historic moments are, there is still risk of similar backlash. If a mistake is made and incorrect information is published, to later be discovered, or if harmful defamation were to occur, the community can be ruthless. Whilst digital preservation may be overlooked and dismissed, or not fully understood, these services will be highly desired in the future and will therefore be on the radar of publicists.

## 3.9 Benefits Evaluation

Evaluating the benefits of discovering sensitive data is not simple, mainly because many investigations conducted may not reveal useful results, but an idea can be formed of the potential this process has. A digital forensic analysis based on evidence found on a known

[1] As of March 2019, James Gunn has been reinstated to direct Guardians of the Galaxy 3 after careful review and tremendous support from the cast and crew. This still serves as a reminder that in today's society, consequences are applied first, and questions come later. Allegations alone can ruin one's reputation, regardless of the truth or how they have reformed. Once a mistake is made, it is very hard to come back from.

criminal has not been conducted; that is not the goal here. The goal is to enhance digital preservation; therefore, the aim has not been to find specific data of a criminal nature, but to evaluate the potential of the data discovered. This does not mean it should be overlooked as any evidence found relating to a crime should be reported and may in fact change the nature of a collection.

The benefits may seem irrelevant for some of the Australian institutions that do not collect a large amount of born-digital material from donations or other sources. This will however change as the increase of born-digital data will continue to grow, therefore, the need for digital preservation and the processes involved will need to grow with it.

The benefits can be seen and understood from both real-world examples and hypothetical scenarios based on existing use cases. The amount of data retained on a system is substantial; the amount of it retrieved with the aforementioned digital forensic software is abundant. With this, both positive and negative impacts may occur. The mishandling of sensitive data and the surrounding risks can have significant negative impact. However, the discovery of new information may strengthen a collection rather than negatively affect it. Therefore, there is an increase in threat as larger quantities of born-digital data are ingested, which in turn, increases the benefits of the sensitive data discovery capabilities only obtainable through digital forensics.

Sensitive data is a crucial component, not to be taken lightly. Consider documenting an iconic figure, presenting him or her in a particular stature, publishing facts that create a perspective in the eyes of the public, only to have been incorrect. The data discovered through the digital forensic investigations conducted revealed much about the users and their online behaviour and a lot can be determined based on this information. This of course is not simply achieved by having a digital forensic tool gather the data, one must analyse and correlate the findings to establish the information. By investigating the online behaviour, documents, emails, etc., one can discover information previously unknown. Extensive analysis and machine learning can be applied to provide a more comprehensive review of data, but as the investigations have shown, basic analysis and correlation is also able to reveal information about an individual.

Collection institutions are not always going to be dealing with a complete hard drive and may in fact be dealing with a set of floppy disks or other physical media. This does not mean sensitive information is not present as it has been shown in this study how data, even small volumes, can still hold an abundant amount of information. The volume of data and the size of files are not the only determining factor, it is how the media on which the data were located is used and how the data were created that impact the type of discoveries made.

Many more examples can be presented that show large amounts of data extracted from sources of various size and usage. There are many types of sensitive data discoveries that can be made, some of which may be unlikely for some collections at this time. However, due to the reliance on digital technology and the increase in born-digital data, the likelihood of such discoveries is ever-increasing. The more we use digital technology and the novel ways in which we do so, the more data are generated, and in turn, there will be more data for which the collection institutions will become responsible. Records of historical data have always been kept and historical figures scrutinised. One day, everything known today will fall under that category. Accuracy and completeness are objectives every collection institution must aim to achieve and without the appropriate discovery and handling of sensitive data, the risk of failing to meet these aims increases.

# 4 AUSTRALIAN LAW IMPLICATIONS

This chapter contains material previously published in "Australian Law Implications on Digital Preservation" co-authored with Denise de Vries and Carl Mooney, presented at iPres 2019, Amsterdam (Hart et al., 2019).

Collection institutions (Libraries, Archives, Galleries, and Museums) are responsible for storing and preserving large amounts of digital data, which can range from historical and public figure records to state or countrywide events. The ingest process often requires sifting through large amounts of data which may not always be sorted or categorised from the source or donor. It is possible to discover information that was not intended to be disclosed should the donor not be privy to the existence of said material. This issue is typically handled by communicating with the donor. If they have no relation to what has been uncovered in the data, further steps may need to be taken. If the data belong to or are about someone living, that person may need to be contacted, depending on the nature of the data discovered. If the person of interest is no longer living, legally there would no issue disclosing all information uncovered. Implications on living relatives must be considered should the disclosed information be potentially revealing or harmful to them. This can include hereditary health issues, political or religious views, and other sensitive information.

There are significantly more variables to consider, such as public interest and defamation which can heavily impact the decision process following the discovery of sensitive data, all whilst guided, but not necessarily enforced by Australian law. This remains somewhat of a grey area as the entities handling such data are often exempt from these laws and principles, making these decisions ethically and morally based more so than legally. The laws and policies that surround privacy issues, defamation, and data relating to Aboriginal and Torres Strait Islander people and culture are explored. The aim is to raise awareness on potential issues that may arise in collection institutions as well as potential threats already sitting in storage and the laws and policies that may serve as guidelines to help overcome and mitigate such issues.

## 4.1 Privacy Act and Principles - Exemption

Digital preservation is not something that has been standardised among the many institutions performing such actions. Some institutions are progressive and are actively making advancements in digital preservation, whereas others are still in their infancy when it comes to preserving digital content. Whilst digitisation of hard-copy material is a role of collection institutions, there are far more potential issues surrounding born-digital content.

The main issues where Australian law may hinder the preservation process are present during the ingest phase and the storage phase, specifically where access is made available. Ingest, storage, and access are being referred to as phases to address all the functions, actions, and on-going activities that encompass them.

Ingest is regarded as the phase which entails the interactions with donors, evaluation of materials, and any processing involved in ingesting materials to the collection.

Collection items may pass through multiple iterations of storage, as there may be different levels of secure storage for certain materials. There are also maintenance and other archival duties involved with storage, therefore, it is not addressed as single function within this research.

The ability to provide meaningful ongoing access to collection items is one of the main goals of digital preservation. There are many different methods for achieving this access and providing it to the public, often differing on case-by-case basis. This may be considered part of the storage phase.

Examples of such issues will be discussed in 4.2.1 Ingest Scenarios.

Whilst not always obligatory, laws and policies exist for good reason. Currently, the main entities performing digital preservation within Australia fall into this area, namely Libraries, Archives, Museums, and Universities. The material these entities store and make publicly available are exempt from the Privacy Act Australia, the Privacy Act 1988 (Australian Government, 2018) and the Australian Privacy Principles (APPs) within. As stated in the National Library of Australia privacy policy:

> *"This policy sets out how the National Library of Australia (the Library) approaches and manages the Australian Privacy Principles (APPs) contained in Schedule 1 to the Privacy Act 1988 (Privacy Act).*
>
> *The Privacy Act regulates how Commonwealth agencies such as the Library collect, store, use and disclose personal information, and how individuals can access or correct personal information the Library holds. It requires the Library to comply with the APPs and take reasonable steps to implement practices, procedures and systems to protect personal information.*

> *The Privacy Act does not apply to library material held, managed and made accessible by the Library, whether published (such as books, journals, newspapers and websites) or unpublished (oral history interviews, photographs and archival collections).”* (NLA, 2018)

The Privacy Act still applies to any user information collected from library services as well as all library employees.

The Freedom of Information Act 1982 (Australian Government, 2022), “Section 13 Documents in certain institutions” states that National Libraries, Archives, and Museum collection documents are not deemed documents of agency under the act, unless categories under subsections 3 and 4. The National Library of Australia does provide a disclosure log, as required by section 11C, and accepts requests to release FOI documents whilst maintaining privacy considerations and not disclosing personal information about people, businesses, and commercial, financial, or professional affairs (NLA, 2021).

The records held regarding our Aboriginal and Torres Strait Islander people have their own surrounding issues along with protocols to guide collection institutions through them. Extra care must be taken in order to maintain the customs of Indigenous peoples and to ensure the handling of their material is done according to their cultural needs. One must again emphasise, there may not be a definitive law regarding such actions, but collection institutions should feel ethically obliged to follow relevant protocols to comply with best practice. Being aware of existing laws and the issues which they aim to prevent, is a necessity for not only adopting best practice, but also preparing for any future changes to privacy law.

## 4.2 Sensitive and Identifying Information

Before understanding the laws that may affect digital preservation, it is important to understand the source of the issues and where they may arise. This understanding is crucial, as quite often the solution must come down to a judgement call, basing decisions on variables guided, but not often enforced, by Australian Law.

Material is often donated to collection institutions, and this can lead to a range of issues. Libraries offer donor agreements which form a contractual agreement between library and donor, stipulating all conditions from both parties and how to handle the data once collected. These agreements may also pass ownership from the donor, removing them from any further say in the matter.

One issue is the discovery of sensitive data. In most scenarios, the donor agreement will typically have instructions in place on how to handle them. However, there are scenarios where the solution is not so easily solved. Firstly, what data are classified as sensitive must be established along with what information can be used to identify an individual. In Part II-Division 1 of the Privacy Act (Office of Parliamentary Counsel, 2017), identifying information and sensitive information are defined as seen in Table 1.

> *"Repositories need to know when a collection offered contains legally protected private files, such as confidential government files; medical records; legal case files; or other kinds of sensitive information, such as Social Security and credit card numbers, whenever possible."* (Redwine et al., 2013)

**Table 1 - Sensitive Data Types**

| Identifying information | Sensitive information |
|---|---|
| Full name | Racial or Ethnic origin |
| Alias/Previous name | Political Opinions / Membership association |
| Date of birth | Religious beliefs/affiliations |
| Sex | Philosophical beliefs |
| Last known address (including 2 previous) | Membership of professional/trade association or union |
| Name of current/past employer | Sexual orientation or practices |
| Driver's license | Criminal record |
|  | Health / Genetic information |
|  | Biometric information / templates |

Regarding the list of sensitive information, these data can be derived by online activity and how the user in question went about their daily activities on the device the donated material was created on. Whilst there may not be an individual element that clearly specifies any of these elements, there may be definitive clues. Much of this information lies deep in a system, obscure, and difficult, if not impossible to find by manual means (navigating directories without the assistance of a digital forensic tool).

One tool that suits this need is Bulk_extractor (Garfinkel, 2013) which can be used to discover anywhere between thousands to millions lines of data deemed sensitive or personal.

With this tool, online activity such as websites visited, which elements within that website were viewed, and any sub-pages visited are revealed. Emails, Facebook, web browser searches, and much more can be derived and analysed to establish information about the user.

For example, health and genetic information could potentially be established if the user frequently researched and visited websites on a health issue. Personal information could be revealed in emails. Religious beliefs and affiliations could also be revealed by online activity, contacts, and communications. Sexual orientation and practices are easily revealed should the user frequent pornographic websites. There is an abundant amount of data that is collected over time, a digital footprint, something which the average user typically will not put much effort into hiding. This data has much potential, both good and bad.

The following is a real-world example. The data have been taken from a real hard drive and processed through Bulk_extractor. Any personal information has been redacted and the example has been carefully selected.

Bulk_extractor detected a high number of URL searchers relating to job seeking:

> "employsa.asn.au", "Job+Search", "Retail+Jobs", "resumes"

Another discovery was visits to the McDonald's login page. It was also discovered that there was an official McDonald's email address assigned to the user. All this information together strongly suggests the user was employed by McDonald's. Correlated against other information and further investigation, it would not be farfetched to say one could establish which workplace the user was assigned to and how much further the investigation could go.

This example shows how individual elements, typically undetectable without the aid of digital forensic tools, combined with other data can reveal a lot about an individual, often sensitive and personal in nature.

Note that whilst numerous records handled by collection institutions are historic and often relating to a deceased person, their sensitive information may still affect any living family members. This relates to health and genetic information. If the information collected indicates the deceased person had a medical condition that is inheritable, this reveals possible health information for their descendants (OAIC, 2018a).

Whilst collection institutions must abide by the laws surrounding privacy with the consumer data they hold, e.g., account information for library users and staff, the collection material itself is exempt from such law. However, this does not mean the laws should not be at least considered as guidelines, influencing policies and procedures for handling sensitive data within collection institutions. The State Library of NSW provided a "Sensitive Collections Material Policy" in 2017 that addresses this with the opening statement as follows:

> *"As part of the Library's collections there is a significant number of records containing people's personal information or, content that is considered culturally sensitive to Indigenous Australian peoples. Examples of these records include medical records, records of children in care, legal records and Indigenous cultural material. Library collection material is exempt from both the Privacy and Personal Information Project Act (1998) and Health Records and Information Privacy Act (2002), however in the spirit of this legislation and based on best practice considerations, the Library sees an ethical obligation to protect people's personal and cultural information. Of equal importance to the Library is enabling individuals to seamlessly access information about themselves and their cultural heritage, especially those who have experienced institutional or other out-of-home care. In light of both of these considerations, this Policy outlines access guidelines to sensitive and private records held in the Library's collections"* (SLNSW, 2017)

The policy goes on to address all instances of sensitive information and lists time restraints for each type of record. Using the privacy laws as guidelines for ethical obligations is something more collection institutions should aim for as it provides a more trustworthy repository for people to commit to and prepares that institution for any future legal changes.

### 4.2.1 Ingest Scenarios

One of the key elements that must be identified is how the donated material relates to the donor and how they came into possession of it. There are many possibilities which change the severity of risk associated with handling such material.

Example 1 – The donated material belongs to and is data about the donor

Example 2 – The donated material is of ancestry significance to the donor

Example 3 – The donor has no relation and has discovered or purchased media in which the donated material was found (known material of significance to collection institution)

These examples relate to events prior to ingest as they would dictate how the donor agreement is written up. However, once the data have been collected and processed, further issues may arise as information can be discovered that was not intended nor covered specifically in the donor agreement. Even if the donor had searched through the material before handing it over, there is a chance they missed something. With training and the right tools, significant amounts of information can be uncovered on a system in obscure places, as well as rich amounts of metadata.

Following Example 1, once the donated material has been analysed, should sensitive information be discovered, further decisions must be made based on what the sensitive information is. If this is covered in the donor agreement, then action should proceed as stated within the agreement. If the agreement does not cover the discovered data and the donor is available, they would need to be involved with any decisions on how to proceed with the uncovered material. There are a few more variables that complicate this procedure. The information may incriminate the donor and depending on the severity and nature of the discovery, law enforcement may need to be involved.

If this scenario was based on Example 2, this may lead to difficulties for living descendants, however, if no direct harm is caused by disclosing the information, legally there is nothing preventing it. The descendants may fight it and they may try to sue for defamation on behalf of their ancestor, or themselves. However, it should be noted that this is a grey area with an inconsistent history.

Another outcome, more likely to occur with Example 3, is the information discovered on donated material may be withheld from the public in their best interest. This may be relating to a public figure, loved, and idolised by the country where the discovered material, whilst harmless, may alter how the public sees that figure. Alternatively, the information may need to be disclosed in the best interest of the public, commonly known as "Public Interest Disclosure" (Queensland Ombudsman, 2017). The donor would have likely signed all ownership of the material over to the collection institution as it has no relevance to them, meaning no further involvement from the donor is necessary in any decision-making. There may be policies in place that help in handling such situations, but for many smaller institutions, this may be unprecedented which ultimately makes this an ethical and moral decision.

It is situations like these that make this field difficult to develop definitive solutions for because no two cases will be the same, there are always grey areas and variables that complicate decision-making. Therefore, it is suggested that collection institutions that may not be equipped to handle these events look to the experience of others, use relevant law as guidelines, and always consider who the data can affect.

### 4.3 Laws

As mentioned, collection institutions such as national and state libraries and archives are exempt from privacy law regarding their collection material. However, it is important to familiarise oneself with the Privacy Act and the APPs to determine if one is classified as an APP Entity. The APP guidelines define an APP Entity to be an organisation or agency. The APP (OAIC, 2018a) define an organisation to be:

*an individual, a body corporate, a partnership, an unincorporated association, or a trust. This excludes organisations such as a small business operator, registered political party, state or territory authority, or a prescribed instrumentality of a state.*

The APP defines Agencies as (but does not include State or Territory agencies):

*a minister, a department, a federal court, Australian Federal Police, a Norfolk Island agency, the nominated AGHS company, an eligible hearing service provider, or a service operator under the Healthcare Identifiers Act 2010. Individuals may also fall under the agency category if they hold or perform duties of an office established by or under a Commonwealth enactment, or duties for the Governor-General, a Minister, as well as bodies established or appointed by them.*

The APPs outline how personal information is handled, used, and managed by APP entities. This applies to most Australian and Norfolk Island Government agencies, private sector, and not-for-profit organisations (with an annual turnover greater than $3 million), private health service providers, and some small businesses. Small businesses ($3 million or under) have responsibilities under the act if any of the following are true:

*Private sector health service providers, sell or purchase personal information, credit reporting bodies, contracted service providers for a Commonwealth contract, employee associations registered or recognised under the Fair Work Act 2009, opted-in to the Privacy Act, relations to another business covered by the Act, or prescribed by the Privacy Regulation 2013* (OAIC, n.d.).

Both the Privacy Act and the APPs are quite extensive, so each principle will not be discussed in detail, but the following is a list of the 13 APPs from the Privacy Act 1988, Schedule 1:

- APP 1—open and transparent management of personal information
- APP 2—anonymity and pseudonymity
- APP 3—collection of solicited personal information
- APP 4—dealing with unsolicited personal information
- APP 5—notification of the collection of personal information
- APP 6—use or disclosure of personal information
- APP 7—direct marketing
- APP 8—cross-border disclosure of personal information
- APP 9—adoption, use or disclosure of government related identifiers
- APP 10—quality of personal information
- APP 11—security of personal information
- APP 12—access to personal information
- APP 13—correction of personal information

Data security and privacy is always a current issue, ever changing, and highly desired. New Government Legislation Acts and policies are often being created, as are current ones being reviewed and amended as needed. Therefore, it is beneficial to be aware of such changes, for they may not be obligatory at the present time, but things can change.

The European General Data Protection Regulation (GDPR) is a prime example as many would be aware by the policy updates from each service subscribed to. All Australian business will need to comply if they have dealing in or with the European Union (EU). This includes having a branch in the EU, offering goods and services in the EU, and even if the business is monitoring individuals within the EU. The GDPR shares many requirements with the Privacy Act 1988, but there are new additions that are not covered in the Act, one of which is the right to be forgotten (OAIC, 2018b). Whilst compliance may not be mandatory, careful review of updated polices and requirements can lead to adopting best practices and better policies. Furthermore, the California Consumer Privacy Act of 2018 (CCPA) (Bonta, 2018), effective as of January 2020, is a worthy mention as it is “setting the pace for privacy legislation in several US states” (Loukides et al., 2020). This shows influential change and how policies and legislation can and will change. Being mindful of jurisdictional differences is important, even if Australian public collection institutions are not categorised as a business under this particular act. Remaining aware is the careful strategy should future changes become impactful.

### 4.3.1 Collection Institutions

There are a few circumstances in which collection institutions need to consider law. These include holding information, making it public, and how the information is being used. The main area of focus is the publicising of information, as this is where the biggest potential threat lies. There are also risks surrounding the content held within collection institutions, however, there are restricted sections where this information is kept from the public. These sections would require special access or permissions by the author or representatives. In fact, the National Library of Australia's (NLA) restricted area, known as the 'Secure Room – Restricted'(SRR) is said to be almost as hard to access as a *"bank vault with its door shut"* (Gidney, 2016). Content is held within the SRR for various reasons, some of the main ones according to Gidney, the author of the blog (Gidney, 2016), include:

- Secret/Sacred Indigenous Material
- Litigation – Ongoing court cases/upheld claims (defamation)
- Commercial in confidence
- Pornography
- Refused Classification (RC)
- Publication with significant/dangerous errors

This list alone illustrates the need to carefully consider what information is made public as the potential risks involved could be quite severe should this listed content not be made secure. Secure areas also serve as a holding place for original documents that may have had information omitted for publicly accessible versions. Gidney listed one such case where 1997, *Goodbye Jerusalem* by Bob Ellis had a sentence omitted that made some offensive and damaging claims. Furthermore, on the topic of making information public, the disclosure of information marked "commercial in confidence" is forbidden without permission from the supplier. This includes any information that may result in damages to a party's commercial interests, intellectual property, or trade secrets (Global Negotiator, n.d.).

### 4.3.2 Defamation

Defamation is defined similarly from country to country, but one of the better definitions posted in an article in "the news manual", sourced from the British Defamation Act of 1952 is defined as:

> *"The publication of any false imputation concerning a person, or a member of his family, whether living or dead, by which (a) the reputation of that person is likely to be*

> *injured or (b) he is likely to be injured in his profession or trade or (c) other persons are likely to be induced to shun, avoid, ridicule or despise him.*
>
> *Publication of defamatory matter can be by (a) spoken words or audible sound or (b) words intended to be read by sight or touch or (c) signs, signals, gestures or visible representations, and must be done to a person other than the person defamed."* (Ingram and Henshall, 2019a)

Prior to January 2006, defamation law varied across each state in Australia, but is now covered under the Uniform Defamation Law (Doctor, 2007). Furthermore, there was a distinction between libel and slander prior to the uniform law, however, the distinction was already disregarded in five jurisdictions and the rest of Australia followed with the introduction of the new law (Rolph, 2009).

Regarding organisations and companies having the right to sue for defamation, this was possible under the old act, however, under the uniform law, if the corporation exceeds 10 employees, they cannot sue. This does not include not-for-profit organisations, and it does not include individuals within corporations of 10 or more employees if they are identified in the defamatory publication (Ingram and Henshall, 2019b).

With all that in mind, it may seem unwise to publicise information, however, there are defences against defamation claims and they are quite solid. First and foremost, "truth" is the strongest defence, more so now under the uniform law as public interest is no longer a requirement needed to supplement the truth claim (Huan, 2006; Ingram and Henshall, 2019b). If there is substantial evidence proving the information to be true, the defamation claim will not succeed. Should the claim be won, it may result in actions taken such as in the *Goodbye Jerusalem* case where the defamatory statement was omitted in the public version. The truth remains the strongest defence for collection institutions, but can be made void should "malice" be proven, that is, if the information was published with ill-will or with harmful motives. It should also be noted, that should the published material be based on a deceased person, they cannot legally be represented in a defamatory case, even by family members. This of course can change should the published material cause harm for living family members, but they can only claim defamation on their own behalf, they cannot clear the name of their deceased family member (Ingram and Henshall, 2019a).

The other defences include:

- absolute privilege

- qualified privilege
- honest opinion
- innocent dissemination (unintentional defamation)
- triviality.

For collection institutions, innocent dissemination is possible, but unlikely as items should be carefully reviewed before being published. Triviality may also prove to be a worthy defence, but the other defences are not as relevant. Absolute privilege covers speech in parliament and court proceedings, meaning whatever is said and whatever motive behind it cannot be used to sue for defamation. The reports of these proceedings are then protected by qualified privilege if the report is honest, for the public, or the advancement of education (Ingram and Henshall, 2019b).

## 4.4 Aboriginal and Torres Strait Islander Material

Within Australian collection institutions, historical records are held containing information on Aboriginal and Torres Strait Islander affairs. There are unique policies and procedures for dealing with such records, one of which is commonly used in libraries called the Aboriginal and Torres Strait Islander Library, Information and Resource Network (ATSILIRN). The ATSILIRN protocols act as guidelines for librarians, archives, and all information services that interact with Aboriginal and Torres Strait Islander people or handle materials with such content (ATSILIRN, 2012). The protocols were published in 1995 by the Australian Library and Information Association (ALIA) and were then endorsed by ATSILIRN. Updates to the protocols took place in 2005 and again in 2010, with 2012 being the latest revision. Once again, these serve only as guidelines, they are not definitive and must be interpreted and applied in context for each issue or situation the protocols may be needed. The protocols cover the following categories:

- Governance and management
- Content and perspectives
- Intellectual property
- Accessibility and use
- Description and classification
- Secret and sacred materials
- Offensive
- Staffing
- Developing professional practice
- Awareness of peoples and issues
- Copying and repatriation records
- The digital environment

Due to Indigenous protocol and sensitivities, some Aboriginal and Torres Strait Islander material must be locked in secure sections of collection institutions, an example of which can be found in the SRR of the NLA. Some of this material may also impose access restrictions and can only be accessed via special permissions such as content classified as “secret men’s” or “secret women’s” business, adding further conditional access (Gidney, 2016).

In 2007, the National and State Libraries Australia (NLSA) developed a framework to guide National, State, and Territory libraries on how to approach Aboriginal and Torres Strait Islander library services and collections. However, this was superseded in 2014 with the “National position statement for Aboriginal and Torres Strait Islander library services and collections” (NSLA, 2014). Within the position statement, it is made clear that the following policies/protocols are endorsed: The ATSILIRN, The United Nations Declaration on the Rights of Indigenous Peoples, and The National and State Libraries Australasia Guidelines for Working with Community.

The standards that are promoted within the position statement include:

- Rights to be informed about collections relating to the people (culture, language, heritage). The right to determine access and use of such material.
- Inclusion of Aboriginal and Torres Strait Islander peoples in all decision-making processes at all levels.
- Strategies to increase employment and retention of Aboriginal and Torres Strait Islander staff.
- Strategies to strengthen cultural competency across the workforce, raising awareness and knowledge on issues for Aboriginal and Torres Strait Islander users.
- Strategies to make usable copies of collection material to be returned to the rightful people to support cultural and language maintenance and revitalisation.

In summary, the promoted standards aim to ensure rights are given to the people relating to the content, ensuring they have the rights to decide how content is handled and managed, to give the people a chance to be part of the process and to give back to the communities where possible.

Another important position statement from the NLSA is on Intellectual Property and how it differentiates Indigenous content and non-Indigenous content (NSLA, 2010). The World Intellectual Property Organisation describes how intellectual property is expressed by Indigenous peoples with the following principles:

Intellectual property is handed down, generationally (orally or by imitation). It reflects community cultural and social identify. It consists of characteristic elements of a community's heritage. It can be produced by unknown authors or by communally recognised communities and individuals that have been granted the right, responsibility, or the permissions. It can often be created for spiritual and religious purposes and is something that constantly evolves within the community.

How Australian collection institutions handle Indigenous material and peoples is a good example of the importance of guidelines and protocols. If we analyse the "CARE" principles (GIDA, 2019), these concepts are relatable in non-Indigenous matters.

There are three principles regarding ethics, the "E" in "CARE".

- E1 – For minimizing harm and maximizing benefit
- E2 – For justice
- E3 – For future use

These principles ensure ethical benefits and harm are considered from an Indigenous perspective and that ethical data must be collected in a way that aligns with their rights. Metadata should also acknowledge provenance, purpose, limitations, and obligations in any secondary use, inclusive of consent (GIDA, 2019).

Whilst collections are not bound by definitive law, the affect collected material can have on others must be a consideration, making this about ethically based, best practice decisions. This should be standard for all material and not just that of Aboriginal and Torres Strait Islander content.

## 4.5 Summary

Whilst many institutions are yet to encounter the issues discussed in this chapter, it does not mean the potential for such issues to occur is not already present. Institutions are storing data, making selected content accessible, and giving it no further thought once processed regarding sensitive material. Whilst some processing may be involved before and during ingest to discover such data, as well as having negotiated agreements with donors in the event such material is found, it may not be enough. Manually searching material or even using built in operating system search functions is not enough for the discovery of sensitive data. Tools exist that are freely available, easy to use, and extremely thorough. Tools such as

Bulk_extractor and The Sleuth Kit (Autopsy) (Basis Technology, 2018a) can be introduced into workflows to significantly increase the discovery of sensitive information.

Without a thorough investigation, sensitive information may be sitting in storage that could potentially be problematic. It may even be useful information, important and crucial to a collection, revealing information that was previously unknown. Hypothetically, should a disk image be created from computing system belonging to a historic figure and the collection institution wanted to discover as much about that figure as they could, forensically analysing the system will reveal what could not be seen prior. Hobbies, interests, past-time activities, social groups, and much more could be discovered. Whilst these methods are typically used to discover questionable and illegal content, it can also be used to find useful and beneficial data. Both outcomes should be the objective of every collection institution as they may be holding information crucial to an on-going or previously dismissed criminal investigation, or it may simply reveal fascinating new information about an entity within their collection.

The way Indigenous content and people are treated should be the exemplar of how all content and people should be treated. Whilst the protocols differ from culture to culture, the example should be followed, that is the efforts made to the best of the institutions ability to cover all aspects, all scenarios, and all potential issues within their expertise and awareness. By doing so and by following guidelines, preventive practices can be adopted, rather than dealing with issues as they unfold. Admittedly, issues such as those discussed may never surface, depending on what type of digital material an institution is dealing with. However, it never hurts to be prepared, especially given the future will be primarily digital and with uncertainty on how it is going to change, in turn, changing digital preservation practices.

If the only concern is with the laws that are binding and not those that collection institutions are exempt from, then it limits the potential to see future issues, hidden threats, best practices, and to generally consider what is best for people. There is never a one-size-fits-all solution, every issue is unique, and every guideline must be applied in context. Being aware is the first step to being prepared for any issues or changes in law that may affect collection institutions. Applicable laws have been discussed, emphasising how they may serve as guidelines, furthermore, giving insight into the issues that can arise in collection institutions, providing further awareness of current and future threat potential. One cannot prepare for something one is unaware of and it is much more viable to prevent, than to fix, making awareness something to strive for.

The following chapter addresses the tools used within collection institutions within Australia and the U.S. By investigating these data, information is revealed on the potential to discover and handle the issues discussed. If there is no evidence of any form of digital forensic software, it is a certainty that sensitive data are being missed. The ability to discover this information is also under scrutiny, specifically within Australian institutions and their public data, such as information stored on their websites.

# 5 WORKFLOW TOOLS – DATA GATHERING

Initially, the data gathering on workflows was broad, eventually focused and narrowed down to Australian collection institutions. Initial data analysis revealed how the workflows of select U.S institutions changed over a few years, the tools used between them, and the complexity of workflows. The U.S data were based on the university workflows made publicly available from members of the BitCurator Consortium (BitCuractor Consortium, 2018). Each workflow review was based on how each institution processes data from acquisition through to storage. Whilst the number of workflows reviewed is small, significant data can be extracted from the workflows themselves and by comparing the difference in complexity, processes, and tools used. The workflows were treated in two sets as one set was posted in 2012 and the other set in 2016 from different universities. There was, however, one university present in both datasets, although from different departments. Each set had unique characteristics in design, structure, and flow. The 2012 set was designed top to bottom with less linearity compared to the left to right design of the 2016 set. Visually, the 2016 set differed with colour coding and additional notation.

Visual representation enables the display of what tool is being used and how many institutions are using it overall, providing a comparison between the two sets. The majority of the data were extracted from the workflows themselves, but the gaps in the 2012 set were filled from the workings of Gengenbach, (2012) and the interviews that were conducted with each institution from that study.

The initial findings provided a comparative example to use against data collected from the selected collection institutions within Australia. Each state and national library were contacted and asked to participate by completing a questionnaire. The questionnaire was designed to evaluate the level of digital preservation being performed, the tools used, workflows, and any use and understanding of digital forensics. Not all institutions were willing or able to participate and others ceased participation halfway through, but from what was gathered and the responses from unwilling participants, a solid evaluation could be made. In fact, the missing data and the lack of information able to be provided for certain questions was just as informative. As the focus is on born-digital collections, an institution's geographic position and physical collection size are not indicative of the extent to which they may engage in digital preservation. However, the size of an institution, respective of their state or territory, may be a factor when regarding local and government data volumes.

This chapter is presented in three sections. The first section focuses on the 2012 – 2016 U.S datasets, followed by section two which focuses on the Australian data. The third section provides a second comparison as well as a breakdown of some of the tools discovered deemed worthy of further investigation. This investigation is based around whether a tool is considered good, or there are questions as to why a certain tool is being used when there are better options.

Whilst the solution is to utilise digital forensic software and methods, all tools are evaluated and investigated to potentially reveal information about each institution and their knowledge, awareness, resource limitations, and overall maturity level of digital preservation. For example, if an institution is not utilising any dedicated preservation tools and is cataloguing with spreadsheets, the likeliness of adequate digital forensic methods being part of their workflow is slim. Other factors such as the continuous use of superseded and unsupported tools are considered to potentially reveal resource constraints or lack of growth in preservation needs.

## 5.1 U.S Collection Institutions – Tools

Before understanding why differences occur in the two sets of U.S data, one must understand BitCurator. BitCurator is an environment made up of many different tools that are utilised in digital forensics and digital preservation. Many of these tools can be used standalone, outside of the environment, but this adds more steps to the overall process of digital preservation. BitCurator is tailored to be a more complete solution, allowing users to perform many of the tasks required in one place, rather than having to navigate multiple tools in standalone environments. However, BitCurator requires training and therefore may be more convenient for some institutions to install and use the individual software tools as they require them.

The following study details the adoption of BitCurator over the two sets of workflows which shows its influence when it comes to workflow complexity and design.

For the comparison between the U.S and the Australian datasets, the focus is on the individual tools used as well as the tools and utilised used within BitCurator. Any solutions formed from this study will consider institutional preference and requirements. No solution is proposed that limit the institution to using a specific tool or method and if a feature from BitCurator is recommended, there are standalone alternatives.

It is also important to note that the team behind BitCurator and the environment itself is influenced by digital forensics. The benefit of digital forensics is acknowledged and therefore influences the workflows, the tools used, and the overall process. Therefore, these workflows were selected to be reviewed and to form the baseline for the proposed enhanced workflows as this is the intended goal for Australian institutions.

The forensic influence is apparent in the U.S dataset, but not in the Australian dataset. This forms the main comparison and seems to be the factor that heavily impacts workflows. Furthermore, it is important to realise no two institutions will be the same, nor their workflows. Although the data collected are made up primarily of libraries, with some archives, each institution handles different types of data or may focus on a selection of records. Therefore, evaluation must be focused on the tools used and what they are used for. There may also be outliers (a single tool used only once among the dataset), but these could appear for various reasons such as a tool created inhouse, or the first to use a new package. These were not treated as outliers, but tools such as virus scanners and hardware used to read a media device unique to that institution were treated as such.

### 5.1.1 2012 and 2016 Workflows and Tools

The methodology used to analyse and compare the data from the available workflows was based on the following criteria:

- Tools used
- How many times a tool was used (how many institutions used the same tool)
- Workflow complexity (Average nodes, design, flow).

Note that “tools” in this study, are both hardware and software based.

Given that the data were based on similar institutions, just a few years apart, measuring how the workflows changed between 2012 and 2016 and determining the factors that led to these changes revealed interesting results.

Each node within the workflows was analysed to establish which nodes were processes handled by tools and checking adjacent nodes to determine if said tools were utilised effectively. For example, if a node indicates a sensitive data discovery process and lists an appropriate tool for doing so, the adjacent nodes indicate the level handling of any discoveries.

This process was repeated for each workflow and any tools that were already listed would have its “total uses” incremented. This led to a list of tools with a total number of uses, and it is clearly identifiable as to which institution used which tool. For the sake of consistency, all data analysis and results have been anonymised.

The results from 2016 had an additional variable. Any tools used that were part of the BitCurator environment were listed as such. This was determined to be an important variable as it was assumed to be one of the key factors for the differences between the two sets, evident in the results.

The "total uses" variable was used when analysing both datasets together to indicate how many times a tool has been used over the two periods. It may also be an indication that a tool is still being used in the 2016 set that was previously used in the 2012 set. In any comparisons between the two sets, the "total uses" variable is unique to each set, therefore will be a different value.

The 2012 set is made up of seven institutions and the 2016 set is made up of five. It is not an ideal data pool, but availability is a factor. Such data are often not publicly available as this type of data is practically impossible to determine from an institution's website and available documentation (policies) which were evident in the Australian institution dataset.

The 2012 workflows were designed similarly to flow charts, therefore, determining how many nodes were within each workflow was easier to establish. However, the workflows from 2016 used a different style of modelling, making use of "swimlanes" to indicate different users or systems as well as colour coding and descriptive notation. This led to duplicate nodes which were not counted towards the total node count. Figure 6 is an example of where this exclusion occurred, showing the duplicate nodes between a Student user and the Digital Preservation Archivist. As these nodes were identical, they were only counted once. Whilst these processes may be performed and implemented differently in practice, the focus here is on the workflow itself and must therefore be taken at face value.

### Dataset 1 (U.S 2012)

This dataset is based on seven institutions, named MEM1 through to MEM7 to anonymise them. Even though these data are publicly available, within this study, all institutions have been anonymised to prevent discovery by elimination for those institutions wishing to remain anonymous. Within this dataset, different levels of digital preservation were conducted among the institutions. Some had complex workflows, making use of many tools, whilst others had much simpler workflows and made use of 2 or 3 very basic and common tools. Not every institution is equal, each handles different types of records and materials, thus leading to diversity between them.

The workflows from this dataset, in comparison to the second dataset, seem more complicated. There are more nodes (average) within the workflows, due to descriptions of lower levels of detail. The second dataset, however, uses a higher level of detail, meaning the inner workings of certain software are not described, process by process. The extra level of detail can be useful; however, in this case it makes the workflow models quite complex and more difficult to follow.

Workflows aside, a total of 20 different tools were used across this dataset, 9 of which were used by only 1 institution, whereas the other tools were used by 2 or more. Seeing which tools were used and by how many institutions was helpful as it suggested the tool was of high quality and usefulness. This by no means rules out any tools used by only one institution as there are many variables to consider. A tool could be new and one of the institutions may have decided to be amongst the first to test it. Word of mouth and awareness help other institutions adopt new software. An institution may be dealing with unique material, therefore need the appropriate software, which is why just because a tool is used once, does not mean it is not worthwhile. Other variables include the hardware available at each institution, which makes compatibility and support a factor, this may also include funding, which may dictate what software can be used.

Figure 7 shows each tool used in the 2012 set and the total number of uses. As can be seen, FTK imager (Forensic Toolkit) had the highest use count. This is a common tool for creating disk images, viewing them, and it can also verify the images once created using checksums. All the tool data can be seen in Appendix C – Tool Data.

An example of tools with a single use not being an indication of its usefulness is presented in Figure 7. FRED (Forensic Recovery of Evidence Device) has only one use, but FREDs are known to be quite expensive and perhaps in excess to requirements, dependent on the level of digital forensics used for digital preservation. Few institutions would be able to justify purchasing a FRED as it would not be used to its fullest potential, especially since the same results can be achieved using open-source, standalone tools. Whilst FREDs are effective, other factors may contribute to its low usage across the institutions.

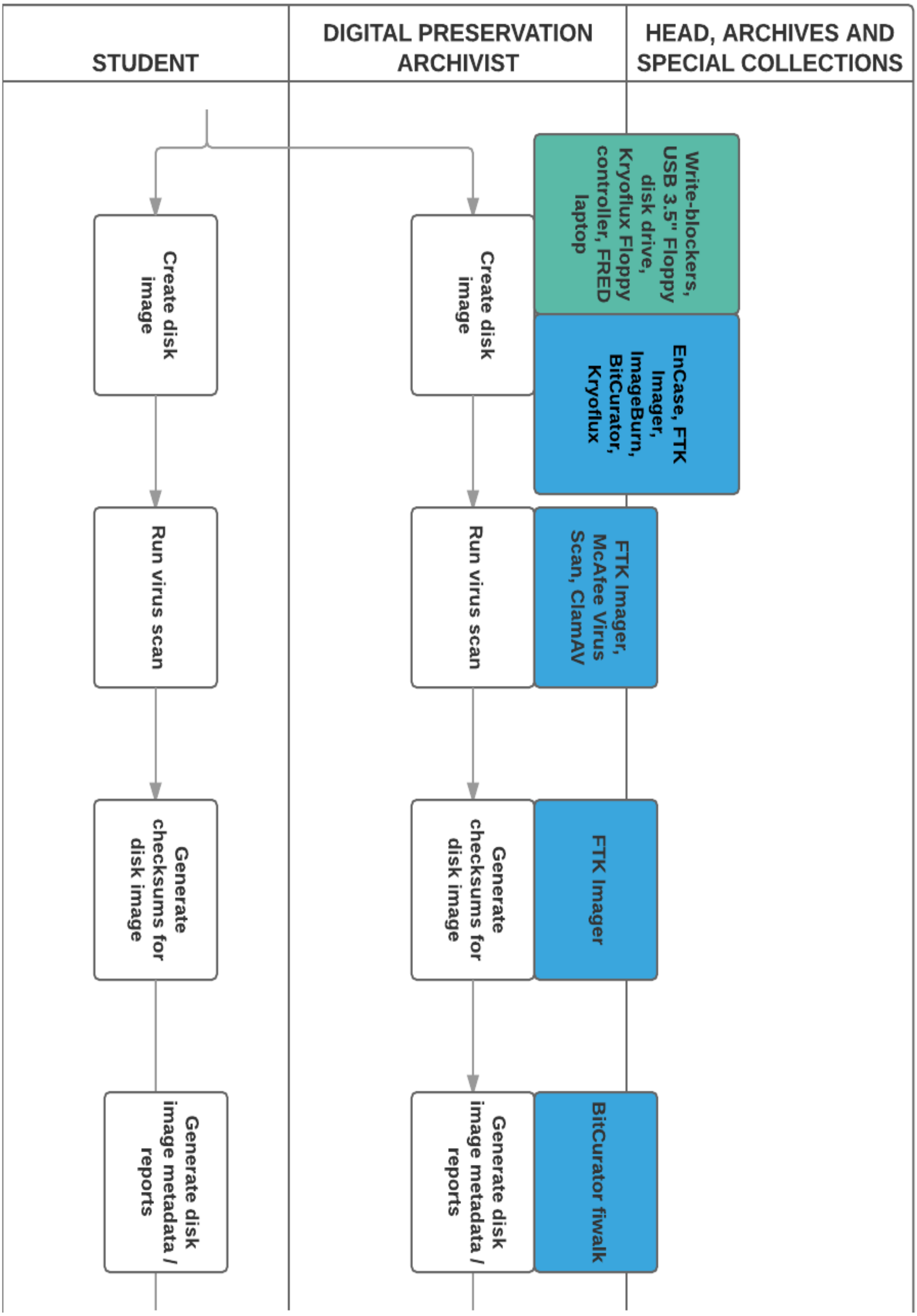

STUDENT
DIGITAL PRESERVATION ARCHIVIST
HEAD, ARCHIVES AND SPECIAL COLLECTIONS
Create disk image
Run virus scan
Generate checksums for disk image
Generate disk image metadata / reports
Create disk image
Run virus scan
Generate checksums for disk image
Generate disk image metadata / reports
Write-blockers, USB 3.5" Floppy disk drive, Kryoflux Floppy controller, FRED laptop
EnCase, FTK Imager, ImageBurn, BitCurator, Kryoflux
FTK Imager, McAfee Virus Scan, ClamAV
FTK Imager
BitCurator fiwalk

Figure 6 - Swimlane Duplicates (Purdue University Workflow Map, BitCurator Consortium)

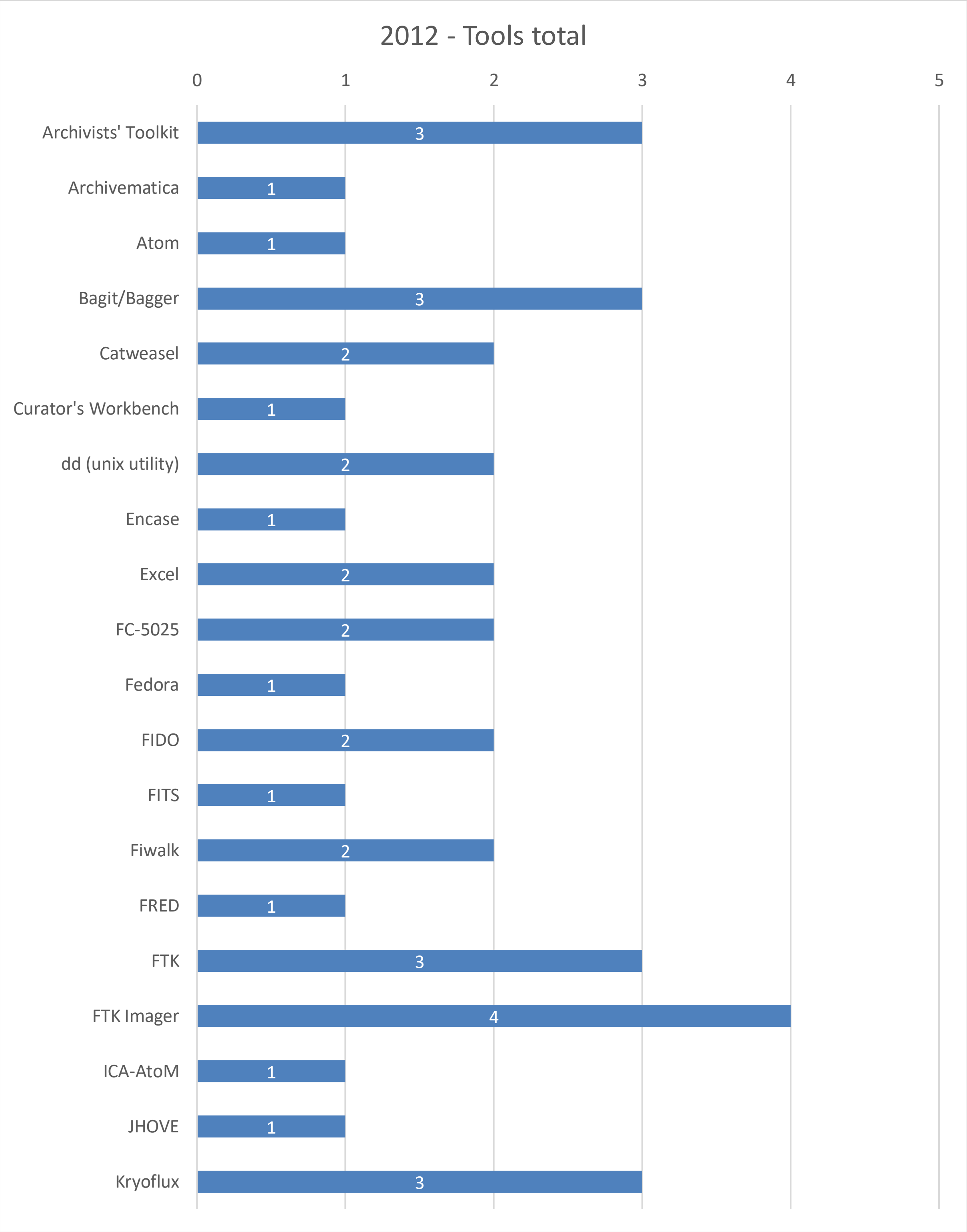


Figure 7 - 2012 Tools Used + Total Count

### Dataset 2 (U.S 2016)

Dataset 2, MEM8 to MEM12, was handled the same way as Dataset 1. Twenty-four different tools were used in this dataset, 13 of which had single uses.

Figure 8 shows each tool used in the 2016 set and the total number of uses:

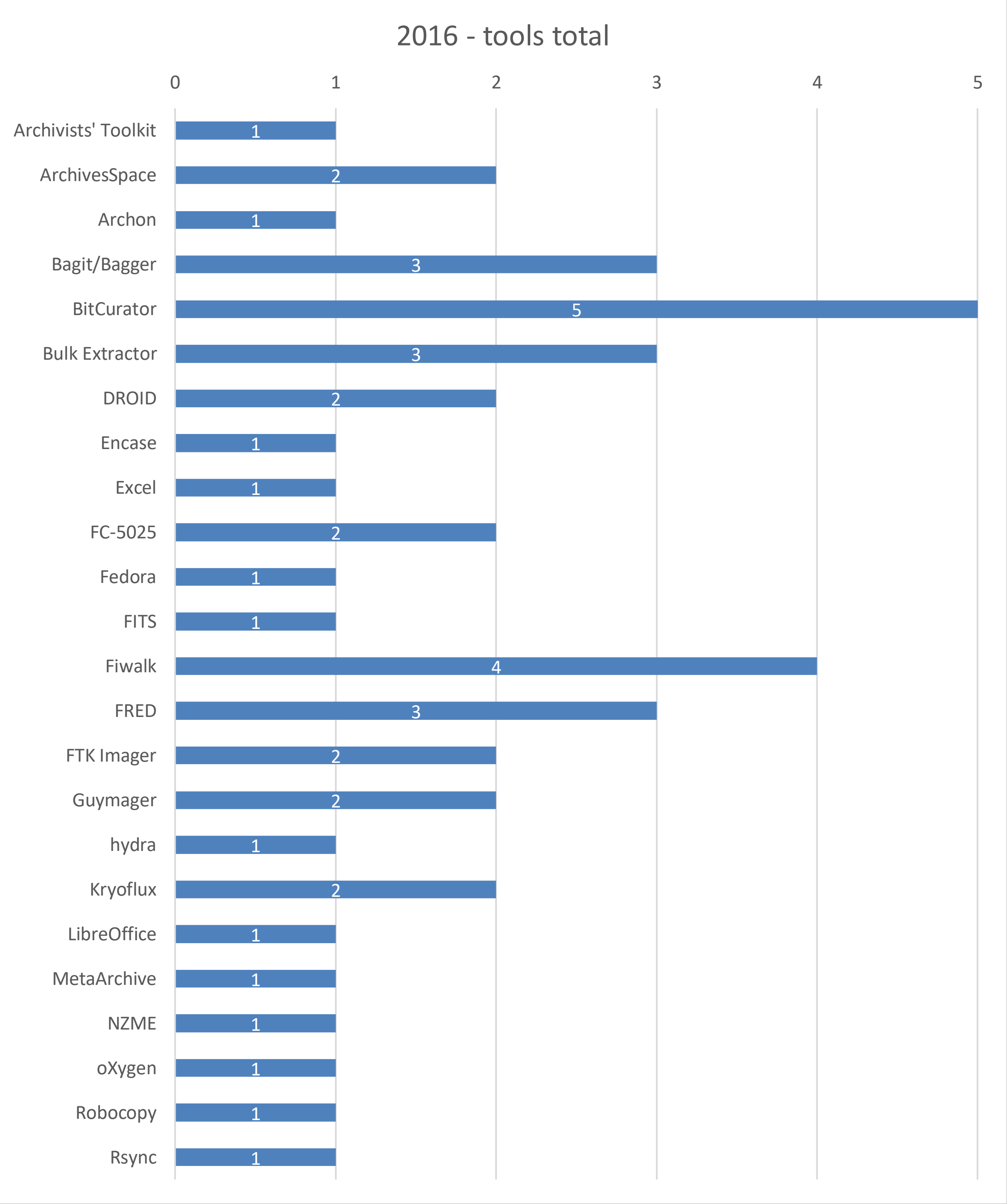


Figure 8 - 2016 Tools Used + Total Count

Dataset 2 has 5 institutions, recording more tools per institution. Dataset 2 had no supporting research; therefore, all the data were extracted straight from the workflow diagrams. This means the true intentions behind using a certain tool are unknown, one can only speculate.

Figure 8 shows each institution had used BitCurator, something that was not present in the first dataset, despite both datasets coming from the BitCurator Consortium. Given that BitCurator was designed for digital preservation purposes, making use of digital forensic tools, the workflows have changed to include steps for the forensic examination of artefacts.

Determining whether a tool is used within BitCurator or through a standalone application is important. This is to ensure that any suggestions derived from this data does not enforce limitation. The workflows visually presented this information.

There were instances of both standalone and BitCurator tools. All instances of Guymager and Bulk_extractor were used within BitCurator, including Fiwalk, in which 3 out of 4 uses were clear. There was 1 use of Fiwalk which remains unknown as to how it was accessed. Many of the standalone tools used may also be used within the BitCurator, which may be determined by preference and training. Sometimes the standalone tool is more efficient to use and somewhat easier if the user is unfamiliar with a Linux based environment such as Ubuntu which makes up the BitCurator environment.

### Comparison

There is a four-year gap between the two datasets, hence there are many variables that could have developed during that period which may directly impact the tools used and workflow design and complexity. As time progresses, the need for digital preservation and its importance is realised as is the volume of data needed to be preserved. One can hope funding has increased resulting in better hardware and software.

This is not an ideal dataset, however, in this area of study, the missing data may reveal gaps where improvements can be made and where processing flaws may exist. The comparison between the two datasets reveals enough information to allow a hypothesis to be formed. This includes identifying tools that are still in use after the four-year gap between the two datasets, as well as seeing what tools are used now that were not being used previously. A larger dataset may have provided additional information such as different tools and usage information on the existing tools. Although there is bias towards BitCurator given the dataset is from the BitCurator Consortium, the data can reveal effective and ineffective tools and workflows.

Effectiveness is evaluated by comparing the tool used and the process for which it is used. For example, if the process of identifying sensitive information is conducted via a tool not known for its credibility in performing this function, such as using the Windows File Explorer search function, one cannot assume the process is conducted effectively. The same can be said when a collection institution is using Excel spreadsheets instead of a dedicated database for archiving and catalogue purposes. Whilst this may work for some, it is far more effective to use tools designed for such functions.

If a tool has been recorded across multiple institutions, it may imply the tool is effective. However, this cannot be assumed based on this alone and must also be evaluated by comparing the process it is being used for.

The 2012 dataset shows more efficiency in tool selection, in that there are fewer single use tools. There are more tools being shared by two or more institutions, potentially indicating effectiveness. However, the 2016 dataset shows better use of digital forensic tools, specialising in the discovery of sensitive data. This has been missing in other workflows, which may be due to the action of sensitive data discovery not being performed or not being visualised and documented well in workflow diagrams.

Effective performance of sensitive data discovery makes use of credible digital forensic tools and methods, ensuring data are processed for analysis before advancing through the workflow. How this information is handled is equally as important, specifically regarding any decision-making based on the discoveries and how the data is stored which includes access permissions.

“When” an institution starts performing digital preservation can also impact its tool selection. If staff have not yet been trained to meet the required experience, they will not be able to use selected tools effectively without prior background knowledge. Without the understanding and implementation of standards, there may be difficulty in selecting tools suited for an institution’s needs which will vary for each institution.

Figure 9 graphs the two datasets together to give an overall view of the tools used. This identifies the tools that are unique to each dataset and the tools are shared between the two. With this information, we can see that certain tools are still in use after the four-year gap between the datasets, giving indication that they are still meeting institutional requirements. However, extending the usage of a tool due to resource limitations or not being aware of updated alternatives are also factors.

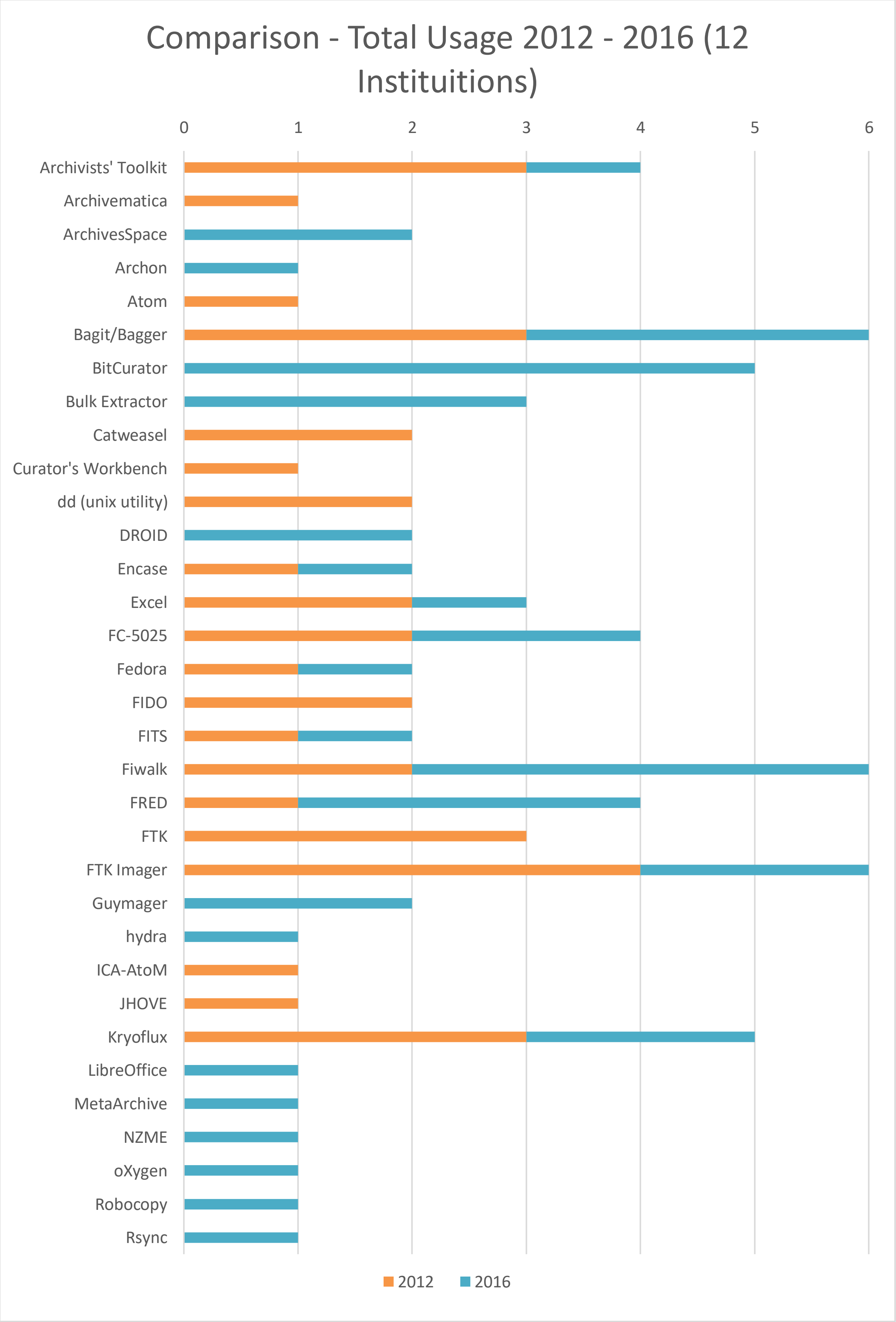


**Figure 9 - Comparison Graph**

Tools that have been discarded as well as new tools that have been adopted in the 2016 dataset are of interest. This information provides a standard selection of tools used across multiple institutions.

The Australian dataset is discussed in the following section. Section 5.2 is presented in an alternative manner as the data gathering method was conducted via both public information and direct communication with participating institutions. A review process was performed on the websites for each State and National library within Australia to assess their publicly available information on policies and other related preservation documentation. The final part of this chapter reveals the tools used within these institutions, derived from the data collected from the questionnaire responses.

## 5.2 Australian Institutions

The current focus of this research is the improvement of Australia wide collection institutions. Every country presents new challenges with jurisdictional laws, policies, and ethical views. The scope of this research was narrowed down to improving digital preservation within Australia by incorporating digital forensic tools and methods. With this incorporation, Australian institutions can perform better sensitive data retrieval and handling.

It is evident that institutions all around the world are at different stages with digital preservation. Some institutions are at a higher level of maturity, whilst others are still in their infancy. This is also true for the institutions within Australia. Whilst the primary collection institutions of focus are libraries, the findings and solutions proposed in later chapters are applicable to all institutions handling born-digital data and performing digital preservation.

### 5.2.1 The Review Process

The reviews were conducted on each state library of Australia, the National library of Australia and National Library of New Zealand. The primary objective was to discover information about each institution’s digital preservation process through workflows and tools. The optimal details would establish processes and procedures from ingest right through to storage and management. Ideally, workflows will indicate what level of digital forensics is being used, whether it is being used effectively, or if there are better solutions available.

Donor agreements and policy availability was also documented, describing how easy it is to find this information, how transparent it is, and where there is room for improvement.

The first step was to review the websites of each library and to identify what information was publicly available, how easy it was to find, and how well it was presented. This helped in determining what information was lacking, therefore, establishing a list of questions to present each institute in the form of a questionnaire. The aim of the questionnaire was to ascertain the awareness of digital forensic tools and methods that may be used for digital preservation processes. Understanding each institution's workflow and their maturity level of digital preservation was important to establish the viability of implementing the suggested enhancements. This was not necessary for the comparative data from the U.S collection as the public information (workflows) display a digital forensic influence, whereas little can be found from public records regarding Australian institutions. Some of the participants chose to have their institution anonymised, therefore, all responses were anonymised by removing names, province, and any unique software that could be used to identify that institution.

The list of criteria used when reviewing each state library's website includes the following:

- **C1** - How easy is it to find information on digital preservation policies?
  - How many mouse clicks/breadcrumbs are required?
  - Are the policies accessible from the home page?
- **C2** - Are all the policies stored in a single location?
- **C3** - Do the policies include: digital preservation, donor agreements, and any supporting documents?
- **C4** - Is the digital preservation policy unique to the institution?
- **C5** - How informative is the policy?
  - Can adequate information be derived about the institutions position with digital preservation as well as their process and workflow? (includes software/hardware usage)
- **C6** - Can donors establish how much control they have over their material from the donor agreement documentation, without communication with the institution?

These criteria were derived from two main concepts. The first focus is based on web design and information structure, which covers how information is provided and the effort required to navigate the website to gain access to this information. Each website was navigated to determine their effectiveness via using link and breadcrumb navigation, sitemaps, and search functions. Website design tips on what to do and what not to do are well established in online articles. For example, navigation and reducing the users effort needed is promoted on the top three online articles about web design "do's and don'ts" (Babich, 2018; Crazy Egg, 2018; Rosa, 2018). Research on usability and website design principles look deeper into these attributes with user testing. The same principles are stressed based on usability, with navigation being emphasised (Faisal et al., 2016; Mvungi and Tossy, 2015; Perdomo et al., 2017).

The second focus was on the information itself and how well the message is conveyed. With this, two principles come to mind.

> *What is the point of having information if it cannot be found?*
>
> *What is the point of finding information if it does not help?*

Websites are changed and updated often. Some information may be outdated, however, the reviews conducted provide an overview of good and bad design regarding information stored on a website as well as how transparency plays an important part.

The main characteristics sought after are:

- Information should be easy to find.
- Information should be grouped and centralised (e.g., all policy information should be discoverable from a single location).
- Breadcrumbs and links should be kept to a minimum (limit the number of clicks required to locate information).
- Information should be complete and effective in conveying what the user requires.
- There should be no broken links or redirections to third party webpages that are no longer available.

Each institution was reviewed under anonymised IDs (from L1 to L10). Although the data collected from the questionnaire was anonymised, the information provided in the following reviews is public knowledge, therefore, any identifications that can be made do not breach any form of agreement. Anonymisation has been applied to all aspects to remain consistent and considerate of the institutions.

**L1**

Upon accessing the state library's website, the policies are not accessible from the home page. They are accessed through the "collections" drop down menu and via "digital collections". The digital preservation policy can then be found on this page. The policy itself is a small three-page document that outlines the scope of the policy, objectives, challenges, principles, and some other information which is expanded on in supporting documents. The supporting documents can be accessed through the digital preservation policy document.

One of the documents, the collection development policy, offers a lot of required information. This document describes in greater detail the policies for each entity the library collects, this entails access, exclusions, intentions, and any unique requirements for a medium. Within this document, born-digital content is addressed, stating the library does not have the resources to preserve such content alone and must work with other libraries and archives to develop policies, strategies, technologies, and standards.

The state library is in partnership with PANDORA (preserving and accessing networked documentary resources of Australia) (NLA, 1999), a project initiated by the National Library with the goal of archiving born-digital content in a form closely resembling the original content. In March 2019, PANDORA became part of the Australian Web Archive within Trove (NLA, 2019).

The policies on donor agreements are clearly listed in the collection development policy document. This stipulates what material will be accepted or declined, as well as legal agreements and other generic requirements. The document states the process of withdrawal without donor approval in the case that the material becomes unmanageable, and the preservation requirements are not within the scope of the collection development policy. The state library seeks approval from the donor, if the donor is unavailable, the donor's descendants are contacted. Should neither of these be met, approval from the Libraries Board will be sought.

L1 have clearly defined polices regarding digital preservation and they meet some of the listed criteria, specifically on the policies themselves, but they are not as easy to find as they should be. The policies show clear understanding of the importance of a sound digital preservation strategy and the need to maintain digital objects in their original format. However, through the publicly available documentation alone, it is not possible to determine the preservation workflow, what software is being used, and any digital forensic methods in place to ensure the policies are met.

**L2**

L2 has no readily available and clearly defined digital preservations policies. There is information regarding what records the library accepts under its "Selection Principles", but this only includes physical media; born-digital content is not covered. The policies that are available, found through a few layers of navigation, are based around physical media as well. There was a page that was accidentally discovered through a Google search and not via any navigation that exists on the website. This page contains a small article that states their collection includes over 270,000 digital items and that they are committed to collection and preservation for future access.

To further explain how this page can only be found via Google search, when on the article page, it shows the navigation breadcrumbs “Home>features>Digital Preservation Week – Born Digital”, however, navigating back leads to an empty page. Trying to navigate back through the URL bar which is “websiteurl/features/Pages/born-digital.aspx” behaves in the same way. Should the text after “/features/” be removed, it leads to the same dead page as before. Removing the page itself “/born-digital.aspx” takes the user to a government department of education page which requires a state education account to login. Therefore, the page can only be accessed through a Google search because it links straight to the born-digital article, bypassing the navigation.

Transparency is important when preserving digital data as donors and users must know the preservation process their data will undergo. Without this information readily available, there is no guarantee any data processed through L2 will maintain its integrity and authenticity.

L2 offers a “Non-Government Records Deposit Agreement” document which covers a typical donor agreement. The agreement is detailed and allows the donor to carefully stipulate how their donated material will be managed and accessed. This includes detailed copyright information which allows the user to assign the material to the “State Archivist” or retain it themselves. Furthermore, the donor can stipulate when the copyright assignment changes to the state archivist which can be after a set date or after certain events such as the donor’s passing. If the donor chooses to retain copyright, there is further selection criteria which specify what can be done with the donated material in terms of making copies, reformatting, and use.

Information regarding born-digital data and digital preservation could not be obtained via the website at the time of this review.

**L3**

L3 offers a website with easy navigation, allowing the user to access the legislation and policies from any page within the website. The user can scroll to the bottom of any page and be presented with the library contact details, opening hours, newsletter information, and quick links to all the main features of the website, including the policies. The policies are all kept on a single page, clearly listed in a table with supporting documentation for each policy. This is a much nicer approach than is seen in other state libraries where each policy is located near its respective subject.

The "Digital Preservation Policy" documentation clearly defines the libraries scope and principles of their preservation strategy. The document states the principles for each milestone in the preservation process from: create and acquire, preserve, store and manage, and access. By reviewing the principles, specifically the "preserve" section, it is clear the library has the correct ideology when it comes to digital preservation. They address the need to avoid obsolescence and degradation by using migration and normalisation. Emulation is considered where normalisation is not possible. The glossary in the appendix of the document defines normalisation as:

> *"The process of transforming a wide range of file formats to a pre-determined set of file formats identified as being more appropriate for long-term preservation."*

The digital preservation policy documentation is further supported by the "Digital Collecting Strategy" document. This policy refers to the strategic objectives when collecting digital material, outlining what type of material is collected as well as any difficulties for specific formats. For example, the difficulty with un-published born-digital material is discussed which includes complexities with access, rights, and management. There is always some degree of difficulty with born-digital material as it is much harder to authenticate and prove the provenance of the material as its digital nature makes it vulnerable.

Information regarding the donation policy and collection acquisition is covered in their respective documents. This includes the collection development policy, supported by the collection acquisition and donations policy documents. The donation policy summarises the criteria in which material is accepted or denied. It contains a link that is supposed to lead the user to the donation page of the library website, but this link leads to a missing page. However, the donation page can be reached by navigating the website. Donors must submit their offers, describing the material in adequate detail and providing their contact details. When an offer is accepted, a declaration must be signed and then the next steps are discussed. One would assume the next discussion would involve how the material should be handled, copyright ownership, and other legal concerns.

L3 has a well thought out repository of legislation and policy documentation that covers a great deal of respective subjects. Having it all in one place offers ease of access. A lot of this information can also be found within their respective website pages, although not as detailed and missing legal specifications. For example, on the acquisition and donations page, what is accepted and not accepted is clearly listed. L3 follows available standards in digital preservation as it refers to the "National and State Libraries of Australasia Principles for Digital Collecting" in the appendix of the digital collecting strategy document. They are also contributing members of PANDORA.

There is little information regarding the processing involved the material is being subjected to during their digital preservation strategy. "What" is being done is stated, but not so much "How" it is being done, which is important information.

**L4**

The L4 website offers a similar initial navigation panel as some of the other websites, having links to important pages at the bottom of each page. However, once the user has selected a high-level subject link, they must continue to follow link after link, page after page, in order to get to the information desired. It takes a minimum of three links to reach a specific policy. Once the user is on a policy page, navigating to others is relatively easy as there is a side navigation column that provides structured headings which then reveal all the sub-categories belonging to it. Whilst the information is not hard to find, it can feel a little tedious and not as seamless as it could be.

The policies themselves are mainly embedded in the webpages with few external documents. The digital preservation policy is primarily made up of definitions and what digital material includes. Furthermore, the policy contents state the legal acts with which the library complies and then discusses their digital object management system (DOMS). DOMS allows efficient storage, management, and access. L4 is committed to abiding by best practice and contributes to PANDORA.

Regarding donations, the standard procedures are taken. An online form must be filled out and then reviewed before a final decision is made. However, unlike most libraries which allow the donor to stipulate conditions, it is stated that when material is accepted, they reserve to right to *"catalogue, store, conserve, and provide access to material at its discretion.".* Furthermore, the library generally does not accept donations offered with conditions, but there are exceptions if the donated material is significant in improving the strength of the library's collection.

No information could be found regarding workflow, procedures, and technical specifications.

### L5

Navigating the L5 website is simple and straightforward. By navigating to the bottom of any page, access to all the major pages is provided, including the policies page. This page lists all the policies in one place, majority of which are external documents. There are other pages discovered through the "collecting L5" page which can easily be accessed from the home page or the site map. Many of these pages link to generic policies and procedures located on the National and State Libraries Australia website, specifically the principle of digital collecting.

The donor agreement form is basic, allowing only for a description of the material and a list for each item within the material. Further stipulations must be discussed with the library if the material is accepted.

There is no specific or unique digital preservation policy for the library. The supporting documentation, like most others, simply states the type of material the library accepts with little information regarding any policies or procedures they have in place. The majority of the reliance is on the national standards and there is no information regarding the library's digital preservation workflow.

### L6

The L6 website offers a wealth of information regarding digital preservation. However, some of the information requires far too many steps to access and does not follow a logical breadcrumb trail. For example, to find how the library processes and preserves digital objects, the breadcrumbs lead as follows: "home > our services for publishers & authors > legal deposit > preserving digital objects". Despite the navigation, the information provided regarding digital preservation is abundant. It is very clear that L6 are quite active with born-digital data and digital preservation, being one of the only libraries to give some insight into their preservation process.

There is a system used by the library, which shall remain unnamed, that is used for long term digital preservation. The system uses checksums throughout the preservation process during several stages, namely MD5, SHA-1, and CRC32. This allows material being processed to be validated at each stage of transition. Once the material reaches permanent storage, annual checksums are preformed to ensure integrity. If there are any issues with material it is noted within a provenance note stored in the metadata for that material, stating any issues or mitigation taken. Each item is stored with a metadata object. The system can also control access to the digital objects and is carefully handled by a handful of staff members.

There are two external policy documents that relate to digital preservation, one being the preservation policy and the other the conservation standards.  These documents contain highly detailed information regarding the environment in which digital content is stored. This information includes specific details about air quality, heat, and lighting.

Regarding donations, there is no formal documentation or form available through the website, instead the processed is handled via communications through email. Donors are asked to email the library with all the information they can discover about their material and are encouraged to send samples of their data. The library will also help in this regard if the donor does not know enough information about their material.

So far, in the order of which each library has been reviewed, L6 is leading regarding digital preservation and the information they offer. However, although the information provided is informative and gives some insight into their preservation workflow, there is still not enough transparency to get a clear understanding of the underlying technical details. This includes exactly what software is used, how it is used, what processes are manually handled, and which are automated. This information is necessary when determining the accuracy of a workflow and where digital forensics is being utilised or where it could be implemented or improved.

**L7**

L7's website does not follow the conventional design that most of the other state libraries follow. Little can be said about the website or the information it contains regarding policies and procedures. Digital preservation is a current goal in the library's strategic plan, but in its current state, there is no useful information on the matter.

Regarding donor agreements, the library offers a detailed submission form along with a guidelines document. The agreement form allows the donor to carefully list all stipulations they may have with their material, including the procedure should unwanted material be discovered in the donor's collection.

### L8

L8 offers a vibrant website with lots to look at. It does follow conventional navigation, offering a panel of links to access each page with easily navigable breadcrumbs as well as a navigation source at the bottom of each page.

The donor agreement is handled through an online form that allows a user to submit their application for donation. Should the library agree to accept the material, further documentation must be completed which is quite detailed, allowing the donor to specifically outline all stipulations and conditions.

The digital preservation policy is the standard policy found in most of the institutions, with no unique policies to the library itself. No technical information could be found.

### L9

There was little to review for L9. Most policies link back to NSLA or are dead links. Regarding donor agreements, donations are only accepted in exceptional circumstances and the items must be of local historical significance or gifts from other governments or organisations.

### L10

L10 manages a basic website that is easy to navigate using breadcrumbs and links, allowing information to be easily discovered.

The library offers a lot of information regarding digital preservation which includes the aims of library, scope, goals, and a general explanation of how certain preservation processes are handled. The digital preservation policy the library follows clearly aligns with best practice and common standards seen in other institutions. The policy is quite detailed and covers majority of the preservation process including the challenges and how risks are mitigated.

There is no specific technical information available. What the library does can be seen, followed by a brief explanation on how, but all technical details are neglected. From what is presented, it is possible to establish a basic high-level workflow, but the technical aspects of the workflow are needed to determine if any digital forensic tools and methods are being used, how they have been implemented, and the level of effectiveness.

Regarding donations, the library offers a form for donors to list their material and cover all the necessary criteria. It is rather basic but allows the donor to freely comment regarding any conditions they may have. The library is willing to negotiate when it comes to donor conditions, should they be realistic and necessary.

**National and State Libraries Australia**

The NSLA is a collaboration made up of each national and state library reviewed. New Zealand was partnered under the NSLA, but left in July 2018, which resulted in the National and State Libraries Australasia becoming the National and State Libraries Australia. Although New Zealand are no longer formal members, the NSLA will continue to maintain the strong ties developed through working together (NSLA, 2018).

From this website, the user is provided access to recent and current events, projects, and important news regarding the latest topics. All partnered libraries can be accessed from this website.

Accessing the policies is no trivial task. If one is expecting to navigate the website to find the information they are looking for, they would be much better off using the search bar. Policies are found within their respective subject areas which requires quite a few levels of navigation, resulting in a long list of breadcrumbs. The policies themselves are quite generic and some documentation links to each of the partnered libraries.

For any significant information regarding digital preservation policies, processes, and any technical information, one should browse the documentation of the library they are interested in. However, some libraries link to the NSLA for their policies and do not have dedicated documentation.

### 5.2.2 Review Summary

It is clear each of these libraries is performing or participating in digital preservation in one way or another. Some are performing at a higher maturity, leading the way within Australia and New Zealand, whereas others are still in early adoption. Whilst all these libraries are partnered under the NSLA, there does not seem to be a strong enough standard for each library to follow, in fact, there is noticeable diversity among the libraries in terms of quality and transparency in policies and other related documentation.

Table 2 visually summarises each library and how it met the listed criteria. The criteria used are:

**C1** - How easy is it to find the required information? How many mouse clicks/breadcrumbs are required? Are the policies accessible from the home page?

C1 is based on how easy it is to find information regarding digital preservation policies and donor agreements. This includes how many mouse clicks are required and whether this information can be accessed from the home page, e.g., in one mouse click.

**C2** - Are all the policies stored in a single location?

C2 determines if all the information can be found in one location, e.g., one page with links to all the policies and supporting documents.

**C3** - Do the policies include: digital preservation, donor agreements, and any supporting documents?

C3 identifies if the required information is among the policies, e.g., are there digital preservation policies, donor agreements, and any relative supporting documentation.

**C4** - Is the digital preservation policy unique to the institution?

C4 is determined if the preservation policies are unique to the institution or borrowed from elsewhere such as a generic list of standards.

**C5** - How informative is the policy? Can adequate information be derived about the institutions position with digital preservation as well as their process/workflow? (includes software/hardware usage)

C5 is based on how informative the policies are, do they give enough information?

**C6** - Can donors establish how much control they have over their material from the donor agreement documentation, without communication with the institution?

C6 is determined by the donor agreement, how informative and detailed it is, and whether donors can establish how much control they have over their material without having to contact the institution.

It should be noted, that whilst some institutions excelled in some areas and although some came quite close to meeting the important criteria such as **C5**, none of them had enough transparency to identify a complete digital preservation workflow. Table 2 illustrates how each institution performed according to the listed criteria.

Table 2 - Criteria Matrix

| | **Criteria** | | | 🟩 **Yes** | 🟨 **Partly** | 🟥 **No** |
|---|---|---|---|---|---|---|
| **Institution** | **C1** | **C2** | **C3** | **C4** | **C5** | **C6** |
| **L1** | 🟥 | 🟥 | 🟩 | 🟩 | 🟨 | 🟩 |
| **L2** | 🟥 | 🟥 | 🟨 | 🟥 | 🟥 | 🟩 |
| **L3** | 🟩 | 🟩 | 🟩 | 🟩 | 🟨 | 🟨 |
| **L4** | 🟨 | 🟥 | 🟨 | 🟩 | 🟨 | 🟨 |
| **L5** | 🟩 | 🟩 | 🟩 | 🟥 | 🟥 | 🟥 |
| **L6** | 🟥 | 🟥 | 🟩 | 🟩 | 🟨 | 🟥 |
| **L7** | 🟥 | 🟥 | 🟨 | 🟥 | 🟥 | 🟩 |
| **L8** | 🟥 | 🟩 | 🟩 | 🟨 | 🟨 | 🟩 |
| **L9** | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 |
| **L10** | 🟥 | 🟥 | 🟩 | 🟩 | 🟩 | 🟨 |
| **NSLA** | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 |

Based on the reviews of each institution and the criteria specified, the recommended order of importance, determined by the ease or difficulty in finding the sought-after information and how well it was conveyed, is as follows: **C4, C5, C3, C6, C2, C1**. It is extremely important to have the policies be informative and transparent on how the institution operates regarding digital preservation. Donors have a right to know exactly how their data are going to be handled and cared for. There are various means for digital preservation, there are many methods, some of which are considered best practice, and some methods are better suited for certain material. Without knowing this information, how can one be assured they are donating to the right institution?

Although **C2** and **C1** are listed last in priority order, these should never be overlooked. Information should not be hard to find, and users should not have to spend time and effort navigating a website finding the information they desire, nor should users have to resort to using a sitemap.

Without some form of communication with each state library, it is not possible to establish the required information to visualise the preservation workflows, therefore making it difficult to provide solutions and improvements. This is an expected result as the OPF, (2020) survey results revealed only 22% had a digital preservation policy openly published, with 26% keeping their policies internal. 32.5% were still developing policies and 19.5% had no policy. With these numbers, the lack of transparency is understandable.

Regarding the information that is available, there is much room for improvement and some form of standard for documentation and the transparency of information should be developed for all those under the partnership of the NLSA.

### 5.2.3 Filling the Gap - Questionnaire

The review process enabled a list of questions to be generated that are designed to cover areas regarding donor agreements, ethical standards, digital preservation, and digital forensics. The extracted information is required to support existing information and to discover where information was lacking. For this analysis, a select few questions were chosen from the Appendix B - Questionnaire to establish what tools were being used and to establish different levels of digital preservation within the participating institutions.

The following are the selected questions that provided the data used to establish what tools were being used and for what purpose:

> "What are the common types of born-digital content your institution works with? (File types, documents, Image, video, audio, etc.)"
>
> "When preserving digital content, are there processes involved to add additional metadata (descriptive metadata) to give the digital content context, as well as improving search and retrieval functionality? (this does not include environment or dependency description)"
>
> "Please describe the process and list any tools (hardware/software) used for this process."
>
> "What precautions are in place to ensure digital content is not changed or accidentally modified during ingest and through to storage?"
>
> "What software is used to facilitate the preservation process? (name and version, please)"

> “What is the purpose of the specified software? (E.g., which part of the process does the software facilitate or does it have a unique function?)”
>
> “Please list any forensic hardware and software used: (Primarily forensic software that is typically used for forensic analysis/criminology, but repurposed for born-digital preservation)”

The responses were quite varied; some were detailed, and some were basic. Some of the questions were not applicable to all participants, but majority of those who participated were able to provide satisfactory results. Even when a response was “*not applicable*”, this was useful data. Unfortunately, not all participants were able to maintain communications and did not complete the questionnaire, despite many attempts to reach them.

Each response from the state and national libraries met the requirements for analysis. One archive provided a response that was more complete and informative than the other archives, one of which stating:

> *“As **redacted** is still developing our digital preservation practices and we do not have donor agreements we are unable to adequately complete your questionnaire at this stage”*

All other contact made with archives did not meet acceptable requirements and were therefore omitted, including the one adequate response as this was an outlier compared with the library data.

The data extracted from the responses provided insight into the level of digital preservation being performed in the participant institutions and, based on the software and hardware used, how effective the digital preservation processes may be. Furthermore, how effective some of the processing may be based on the software and hardware used was identified. For example, the diversity in responses to the following questionnaire questions demonstrates the variations and level of digital preservation between two institutions:

Question 2: When preserving digital content, are there processes involved to add additional metadata (descriptive metadata) to give the digital content context, as well as improving search and retrieval functionality? (this does not include environment or dependency description)

> *"Generally no, we do not add additional metadata to the files/ However for at least one instance, for a large transfer, we have added a limited set of metadata at creation of the preservation digital files."*
>
> *"At **Redacted** we use our catalogue record as the "source of truth" for all descriptive metadata. Minimal descriptive metadata is added into our digital preservation system, **Redacted** and embedded into access copies. **Redacted** is where we document mostly technical metadata about the operating system, file system, codecs and other dependencies in the DNX. The minimum of 5 points of DC is added to all files: MATCH POINT, TITLE, AUTHOR/ CREATOR, RIGHTS, and CONTRIBUTOR"*

Question 4: What precautions are in place to ensure digital content is not changed or accidentally modified during ingest and through to storage?

> *"Generally, at this stage we have few ways to ensure that modification doesn't occur. We ingest files in the form they are given to us as our system is not set up for digital preservation. The only thing that we can do is create masters and working copies. However there is no mechanism in place to monitor Master items stored in shared storage. We would not know is these files became corrupt. For special cases we have used fixity as created by the Bagit protocol, LOCKS (2 x Ext HDD, 1 X LTFS ), and scheduled checking of data. If there is a mismatch of checksums we would revert to the copy that have the original matching checksum"*
>
> *"We take the following steps to protect the integrity of the files:*
>
> - *write blockers to protect content when appraising and transferring.*
> - *virus checks for all files (malware included)*
> - *We create checksums for all files.*
> - *For the transfer and checksum we use Bagger (LOC) that utilises the BagIt standard (We validate the bags throughout our workflow)*
> - *Checksums are stored in **Redacted** so that files can be validated over time"*

### 5.2.4 Results – Australia

This section focuses on the results pertaining to the tools used within the Australian institutions. Result data from the questionnaire relevant to workflow efficiency and reliability is explored in Section 7.1 Workflow Evaluation - Australia. It is evident that Australian collection institutions, particularly libraries, are progressing in digital preservation. Some of the institutions at the time of answering the questionnaire were in the process of adopting new tools and phasing out older tools. These tools were included for analysis as seeing which tools are being replaced and which tools are replacing them is an indication of growth.

As can be seen in Figure 10, there are no real standards when it comes to what tools to use. Some of the tools mentioned are barely utilised to their full potential, some were just starting to be adopted, and some tools were adopted as the result of training. Training is a factor that determines what tools are used and used effectively. Without proper training, proficient use of tools may not be possible dependant on the experience of employees. Institutions may not be aware of the existence of such tools, nor know how to use them. This is important as not being aware of tools and solutions will lead to issues such as data being ignored or unnecessarily converted or normalised. This can result in data loss, affecting the original data’s integrity and authenticity.

The types of materials being processed predicates the selection of tools used. For example, processing and preserving audio files requires specific tools and this need can be seen in some of the single-use tools that are specifically used for such media.

The questionnaire results provided data describing all the file types each participating institution works with, including which were most common, and a total count of each file. Correlating this data (common file types for each institution) with the tools used for each institution generates two assumptions:

*One - the types of content the institution deals with dictates the tools used.*

*Two - the tools used, may in fact, dictate the types of content that institution accepts or specialises in.*

Some of the tools listed are hardware based, used to read data on physical media. This may also influence the types of files dealt with as it has been stated by one or more of the participants that the means to access files from certain media are not always available.

Each institution has dealt with at least one of each file type, but the most common types are PDF and images (TIFF, RAW, JPEG), with the exception of one institution where Websites were the second most prominent type of material (9000 records), with PDF being first (28,000 records). The statistics on the numbers of each file type within each institution were not provided. Table 3 - Preferred File Types shows each preferred file type.

The final section in this chapter provides a second comparison that includes both datasets and includes a discussion on a selection of tools.

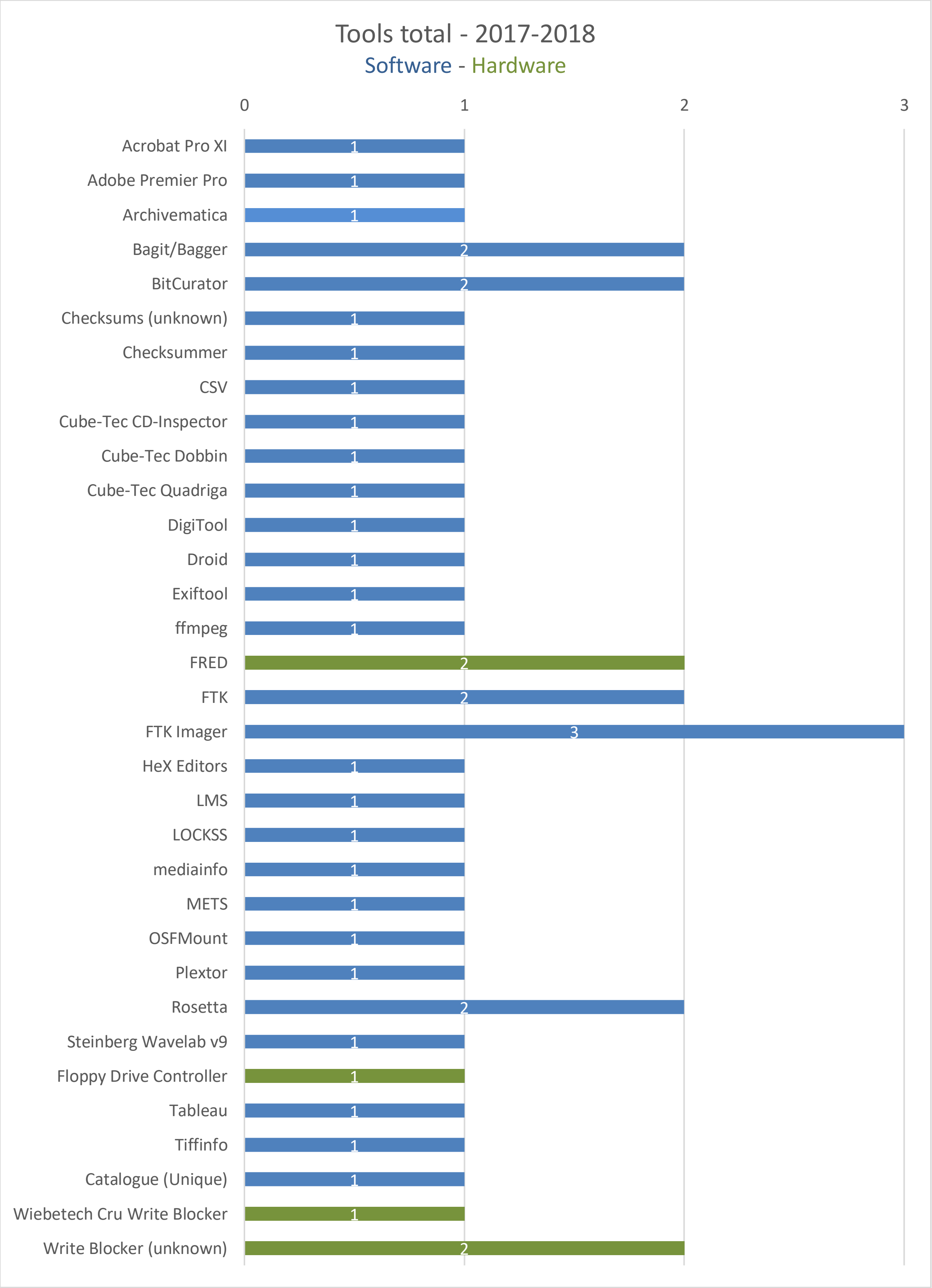


**Figure 10 - Australia Tools Total (6 Institutions)**

Table 3 - Preferred File Types

| File type | Format | | | | | | |
|---|---|---|---|---|---|---|---|
| Image | TIFF | RAW | JPEG | | | | |
| Movie | AVI | MOV | MPEG | MP4 | MXF | QuickTime | |
| Documents | ALL | .doc | PDF | … | | | |
| Transcripts | HTML | JSON | PDF | RTF | SRT | VTT | XML |
| Audio | AAC | BWAV | MP3 | MP4 | Wav | | |
| Web | ARC | | | | | | |
| General | ZIP | .exe | | | | | |

## 5.3 Datasets Combined – Comparison Two

It is evident that the U.S datasets are influenced by digital forensics as they all come from the BitCurator Consortium, whereas there appears to be minimal influence within the Australian dataset. Digital forensics has much to offer digital preservation; however, the benefits of the tools and methods available are not always obvious for collection institutions. The 2016 U.S dataset shows evidence that the workflows actively considered personal and sensitive data by making use of digital forensic tools such as Bulk_extractor.

Issues surrounding sensitive data are problems most cultural heritage institutes archiving born-digital objects are facing currently. Thus sensitive data have been forming the premise and main arguments for this study, as it is something with great potential, both informative and detrimental. It is key for digital forensic investigations (criminal), but potentially overlooked when it comes to digital preservation. The Australian dataset strengthens these suspicions, which is a concern.

The data reveal the 2016 dataset making less use of packaged software such as Archivists Toolkit and there were no recorded uses of FTK, whereas these tools were among the highest recorded count for the 2012 dataset. The addition of BitCurator to the 2016 institutions is a possible cause of this as it serves as an integrated package. Another interesting discovery surrounds the tools with a single use. The 2012 dataset had 9 single-use tools, 5 of which were from 1 institution. The 2016 dataset had 13 single-use tools, 6 of which were from 1 institution, 4 from another, and the rest were spread out across the remaining institutions.

The single use tools from the 2012 dataset include:

- Archivematica
- Atom
- Curator’s Workbench
- Encase
- Fedora
- FITS
- FRED (Hardware)
- ICA-AtoM
- JHOVE

The single use tools from the 2016 dataset include:

- Archivists’ Toolkit
- Archon
- Encase
- Excel
- Fedora
- FITS
- hydra
- LibreOffice
- MetaArchive
- NZME

- oXygen
- Robocopy
- Rsync

The tools listed for both datasets were all standalone. Integrated tool usage consisted of: Bulk_extractor (3 uses), Fiwalk (3 of 4 uses), and Guymager (2 uses) which were used within the BitCurator environment.

The Australian dataset was made up of 26 single-use tools. Some of these were unknown and un-named. There was one account of a checksum tool that was listed as “checksums” and there were two cases of a write blocker in use without any specific information provided. Therefore, instances of checksum and write blocker tools may in fact have more than one use across the institutions.

Two of the six institutions made up majority of the single-use tools, with counts of 7 and 8. Correlation of the data types handled by each institution and the tools they have listed provides a better understanding of the results. For example, one of the two institutions with a high count of single-use tools deals with audio, therefore, specific software and hardware tools are required, often unique and not used by others without a demand for audio work.

The one consistency between both datasets is that even when part of a group with similar goals, there are commonalities and exclusivities. This is unlikely to change as each institution is operating at a different maturity level. Demand plays a pivotal role in this regard as without an increase in preservation needs, an increase in maturity level may not be warranted. This affects the procedures and tools used to accommodate current preservation needs.

Institutions willing to develop and mature are not always averse to adopting and experimenting with new technologies and methods. Other institutions are content with their current state and are waiting to reflect on the results of others before making any significant changes. Tool data and institutional partnerships with varying levels of maturity suggest it is probable that the institutions still in their infancy, with respect to digital preservation, may be influenced by their peers. This would lead to data showing more uses per tool, rather than single-use tools regarding digital preservation processes on standard materials. This relies on institutions communicating and sharing their experience, the tools they use, and the methods they implement. Transparency regarding public documentation on policies, procedures, methods, and tools, will provide examples for other institutions to follow.

### 5.3.1 Tool Breakdown

In this section, some of the tools in the dataset are investigated. The selection of the tools to investigate was determined by the number of times they appear across the three datasets as well as any unique tools that warrant further investigation. Tools with uses and that are not specifically designed for digital preservation, such as the use of spreadsheets, CSVs, and operating system command line functions such as the unix and linux command for data duplication (dd), are not considered. This is not to say tools such as these do not have their merits, nevertheless for large-scale workloads, one would recommend using dedicated tools designed for such tasks.

Figure 11 and Figure 12 are derived from the data containing all the tools used and the counts between the three datasets. The two figures are filtered so that one shows all the tools with more than one use and the other shows all single-use tools. All tools of significance and all tools that are reviewed are referenced in Appendix A – Tool Sources.

The following tools were selected for investigation based on the tools used by the Australian and U.S collection institutions:

- Archivematica
- Archivists' Toolkit (ArchivesSpace)
- Bagit/Bagger
- BitCurator
- Bulk Extractor
- DROID
- Fiwalk
- FRED (Hardware)
- FTK and FTK Imager
- Kryoflux (Hardware)

The selection criteria for each of the listed tools vary per tool. The purpose for each selection was based on frequency across the three datasets.

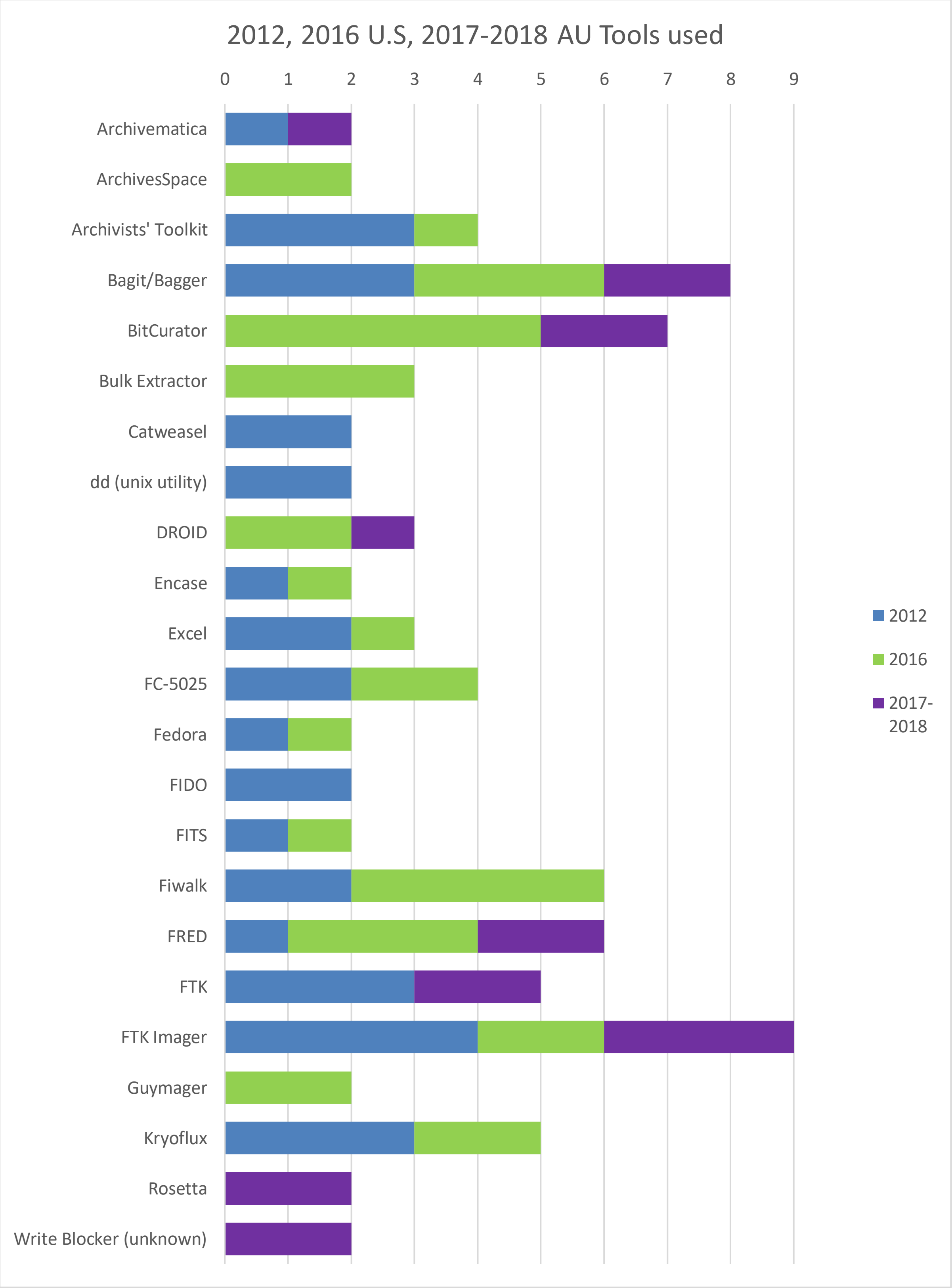


**Figure 11 – Count of Tools Used By 2 or More Institutions**

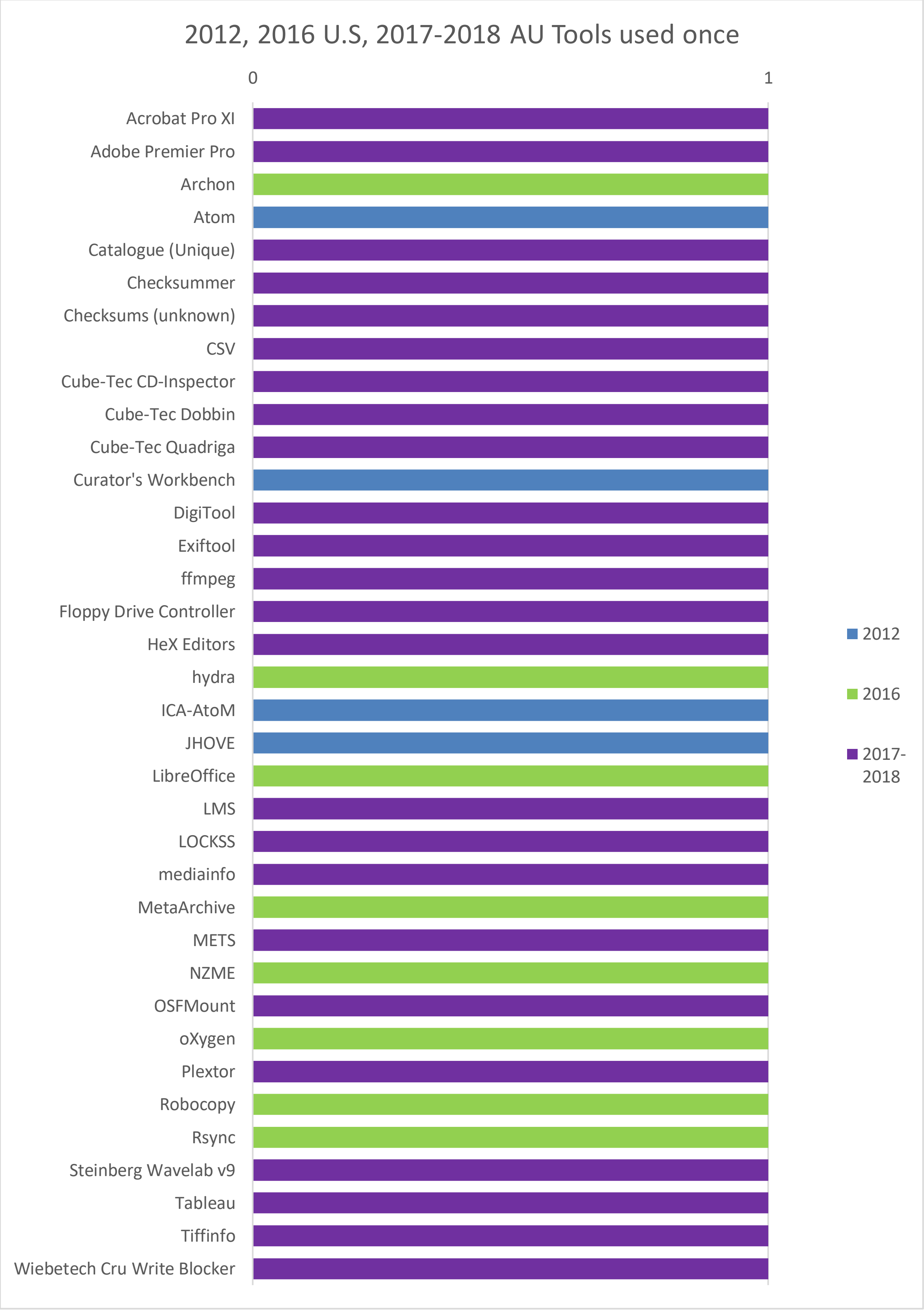


Figure 12 - Tools Used Once

Archivematica was investigated as it is open-source, quite flexible and customisable, and makes use of standards such as METS (Library of Congress, n.d.) , PREMIS (Library of Congress, n.d.), OAIS (CCSDS, 2012), Dublin Core (DCMI, 2018), and also utilises Bagit.

Archivists' Toolkit and Archon were superseded in 2013 with the release of ArchivesSpace 1.0 (ArchivesSpace, 2020). Archivist's Toolkit and Archon both had a single use in the 2016 dataset although having been superseded three years prior. This raises questions regarding the usage of superseded tools and whether there is any risk involved. There may also be a high dependency on the existing tool, where any changes may cause disruption to the workflow as new supportive hardware and software is needed as well as staff training. Implementation considerations would then be required and may take significant time dependant on resource availability. Funding, training, and awareness are factors in this regard.

There are instances where an obsolete tool is still performing adequately. Therefore, the institution will decide against making any changes that have the potential to cause disruption or require additional resources to implement and maintain. This approach, however viable, is not recommended as the risk of obsolescence increases with the continued usage of tools after official support has ended.

The use of discontinued and unsupported hardware and software is not uncommon. A study conducted by Spiceworks, (2019), spanning three years, surveyed 489 "IT decision makers" across North America and Europe, ranging from small, mid-size, and large enterprises from various industries. In 2017, 42% were still using Windows XP on at least one device. As of 2019, this number has decreased to 32%. The last security update for Windows XP was April 2014. This operating system is no longer supported and is now vulnerable to newly discovered security threats that will not be patched. Current and future software updates are not guaranteed to work within the outdated system.

Bagit/Bagger was selected as it was one of the tools with the highest usage and is a file packaging application that shows up in many digital preservation packages and institution workflows.

BitCurator was selected as it is an integrated environment, containing many different tools suited for digital forensics and applicable to digital preservation. BitCurator had a high usage count in the 2016 dataset and two uses in the 2017-18 (Australian) dataset. Although both the 2012 and 2016 workflows were from the BitCurator Consortium, the influence is clear in the 2016 workflows with the presence of BitCurator. The Australian dataset and the responses from the participants indicated BitCurator is not easily adopted without the proper training. One institution from the 2017-18 dataset had training on how to use BitCurator and was therefore able to adopt it and continue to use it effectively. One other institution has BitCurator, but stated it is not often utilised.

Bulk_extractor was only present in the 2016 dataset with three uses. It is a candidate for review as tools purposed for sensitive data retrieval are integral to this study. Tools such as this can extract personal and private data which would be impossible to find by manual searching methods. This is an important step in digital preservation, especially when handling donated materials. Important information may be hidden that could benefit the collection or may be crucial for determining how the institution should proceed with the material in question. Certain discoveries may alter the course of the processing, changing the destination of the material within the collection. It may require redaction or storage in a set location utilised for sensitive material. As the demand for preservation of born-digital artefacts increases, the intake of materials will increase with it, making digital forensic tools more necessary.

DROID – Digital Record Object Identification was chosen for two reasons. Firstly, the file identification features, whilst very useful, are sometimes present in other software and integrated packages. Comparing these two methods of file identification is of interest because whilst integrated packages are convenient, they can be excess to requirements. As DROID only had three uses, two from 2016 and one from 2017-18, determining if file identification is performed or if alternative solutions being utilised are of interest as this can reveal if an institution is effectively, if at all, making use of file identification.

Fiwalk had four uses in the 2016 and two uses in the 2012 datasets. Fiwalk has now been integrated in the SleuthKit and Archivematica and can also be used within BitCurator. Fiwalk had no uses in the 2017-18 dataset and determining what tool replaced this functionality is the primary reason for investigating Fiwalk.

FRED workstations are powerful machines that can safely image multiple drives, conveniently and efficiently. They can be customised and tailored to specific needs and can be significantly upgraded in terms of CPU power. Some institutions make use of FREDs, others have deemed it excessive and unnecessary. The same can be achieved with much simpler and cheaper solutions, especially if the demand for imaging is low. The volume of work is a deciding factor on whether a FRED can be utilised efficiently.

FTK and FTK Imager – FTK imager had the highest use count (combined) out of all the tools and this was an expected result. FTK imager is a tool for creating disk images and viewing disk images. It has built in checksum generation and verification, making it a powerful addition to any preservation workflow. FTK imager is a standalone tool in the Forensic Toolkit (FTK) and there were some recorded uses of FTK in the 2012 and 2017-18 datasets. There were only 2 counts of FTK being used in the 2017-18 dataset, one of which had its usage confirmed in the questionnaire response stating FTK was used in place of BitCurator as it better suited their needs. The second recorded use of FTK is somewhat ambiguous. The institution stated they had recently acquired FTK software at the time, meaning the extent of its use is unknown as are the features of the software to be used. This could have resulted in FTK imager being the only part of the toolkit utilised.

Kryoflux was selected to be discussed from an alternate point of view. It had high usage in the 2012-2016 datasets, but no recorded uses in the 2017-18 dataset. Kryoflux is a hardware solution for reading floppy disks on modern computing systems through a USB and is a proven tool. If a wide range of floppy disk formats needs preserving, Kryoflux is a popular choice. However, as time progresses, the need to preserve floppy disks will diminish.

As of 2020, the results from the OPF, (2020) survey reveal a shift in tools used across various institutions. There are still instances of unique tools with low usage. The top 5 tools amongst the participant list that are being evaluated and tested are: JHOVE, DROID, ExifTool, ImageMagick, veraPDF, and Bagit. Of these, JHOVE and DROID are in production by approximately 50% of institutions, with ExifTool and ImageMagick being in production by approximately 40% of institutions. veraPDF has less than 20% of institutions using it and Bagit has approximately 30% usage. BitCurator comes ninth in the list with less than 20% in production, with just under 40% evaluating and testing it.

These selections and numbers will change and be influenced by the community of which these institutions are part. Various cultures and standards may be developed within these communities which may not be shared outside of the community.

Tools that are built for legacy media may also see a shift as time progresses. For example, the discovery of legacy media may continue for some time; however, tools such as Kryoflux, have already been developed that are fit for this purpose. If these tools are preserved correctly with the appropriate documentation and metadata, continued usage should be possible.

However, should new tools need to be developed due to hardware degradation or damage, the documentation and metadata can help in engineering a modern version of the device.

Obsolescence will affect CDs, DVDs, Blu-ray, and hard drives. Modern personal computers are rarely designed with CD drives and will not contain a floppy disk drive. Solid state drives are continuously advancing, getting faster with more storage capacity, eventually making other hard drive types (PATA, SATA, SCSI) obsolete. This means new solutions to preserving old media are not necessary, but future considerations to current media are. The future of preservation relies on software with less importance on hardware. Hardware will remain a consideration, even solid-state drives have evolved from their initial design to Non-Volatile Memory Express (NVMe) with M.2 form factor which slots into the M.2 slot on compatible motherboards (Kingston, 2020).

As an example, floppy disks became commercially available in 1971, being developed in the late 1960s. They were developed with the current technology at that time in mind and possibly future assumptions. Computing has advanced significantly since then, meaning if a new media were to be developed right now, it would not have the same restrictions and limitations. One could assume a much longer lifespan in terms of current hardware support as technology, whilst continuously getting better, has not gone through significant architectural change in quite some time.

### 5.3.2 Summary

The data presented in this chapter have revealed insights into the current state of digital preservation practice in Australian institutions in comparison to the U.S institutions explored. Data collection for this study was limited as definitive and complete data was not always available due to the different progression levels of digital preservation for each institution. The U.S datasets provided a baseline and some insight into various tools and workflows based on their digital preservation needs. These datasets along with the Australian dataset allowed for comparisons and visual representations of changes and similarities over the span of a few years.

It was discovered that some tools were still being used currently that were being used in the earlier datasets. Some tools were phased out and some were being used that had become obsolete and superseded, whereas others were making use of more modern and supported tools.

The data collected made it possible to establish the following criteria which can determine where the digital preservation process is lacking and could be improved for each institution:

- Which parts of the digital preservation process (workflow) are supported by tools (digital preservation, digital forensics, and re-purposed tools)
- Are the tools used considered adequate?
    - Is the tool being used built for the purpose of its use?
- Are the tools in use current and supported?
- Are the tools used for typical processes found in digital preservation workflows used by other similar institutions?
    - If not, why? (for a unique process or an outdated tool)

**The answers to these questions can identify instances of where tools could be used, but are not, and when they are, if the most appropriate tool is being used. The results from**

to Figure 12 provide insight into these concerns. If there are no recorded uses of a tool dedicated to sensitive data retrieval, how can one be confident that this process is being handled? If only a small number of institutions are utilising such tools, the question remains, how are the other institutions are handling this process, if at all?

If uses of a CSV spreadsheet are identified and the purpose of its usage is typically supported by a dedicated tool, such as database or catalogue, why has such a solution not been adopted? It is questions like these that can help in gauging the maturity level of digital preservation being performed as comparisons to other institutions can be made based on these evaluations.

Based on the criteria specified, this evaluation can be determined by indications of core preservation practices going unsupported by dedicated tools. Other factors considered consist of the usage of outdated tools as well as tools that are not used by any other institutions and why. Processes that could not be conducted effectively without the aid of a tool built for such a purpose are considered a strong indication that the process is not handled appropriately. These issues may arise as a result of resource restrictions, workload, and the mission of each institution. A low maturity rating does not reflect negatively on an institution.

Though the participation rate is low, there is a wide range of values in the responses. Each response came from institutions at various digital preservation maturity levels, from infancy to further developed. The data captured revealed institutions at both the early and later stages of digital preservation implementation, deduced from what tools are being used and for what purposes. This information, with other data from the questionnaire results, validates the assumptions made throughout this study based on the level of digital preservation being performed and the influence of digital forensics on the digital preservation processes.

The next chapter focuses on digital forensics. In Chapter 6, digital forensic tools are examined and reviewed. The usage and features of the selected tools are investigated. The approach of a forensic analysis is only being used where the goals align with those of a collection institution, emphasising data retrieval, specifically sensitive data.

The experiments are conducted on real data, providing real results to be analysed. The output provides an overview of potential benefits and risks associated with sensitive data, displaying the capabilities of digital forensic tools within collection institutions.

# 6 DIGITAL FORENSICS – SENSITIVE DATA

This chapter presents the analysis of data resulting from the investigations conducted using digital forensic software, specifically Bulk_extractor and Autopsy. Bulk_extractor was selected as it is an open-source solution and was present in the 2016 U.S dataset. Personal experience was also a factor, and it is a relatively easy tool to use with minimal training required. Autopsy was selected as it was discovered to be a competitor to propriety forensic suites such as FTK and EnCase. This was a free, open-source solution, a desirable factor for both a student and collection institutions with resource limitations.

The purpose of these investigations was to reveal the amount of data that can be discovered in obscure places, un-reachable by manual methods of searching. The tools and methods explored in this chapter and the data discovered display a range of benefits for collection institutions, from strengthening a collection with new and undiscovered material, to reducing risk of sensitive data being mishandled.

The results presented are only a fraction of what can emerge as the experiments conducted were based on small hard drive sizes and the output was controlled to show only examples of potential. Should the amount of the data increase, the outcome will be significantly larger. The size of the data is not always a determining factor of how much and what data can be collected through these methods. A large hard drive consisting of high-definition movies, considerably larger in file size compared to documentation, has limited potential in this regard. The types of users and how they use the system is reflected in the data gathered and can result in significant differences. For example, a user with high-level computing knowledge will know how to delete data properly and how to reduce their digital footprint, resulting in a different outcome when processing this data. Multiple users on a single system will also lead to additional user data and influence being uncovered.

The primary goal was to see how digital forensic software and processing could enhance a digital preservation workflow. Due to this, the software was not tested to its full potential, as some of the modules are specifically designed to be used in criminal investigations and other forensic purposes. However, the range of benefits discovered offers much to collection institutions, whether it be now or in the future when an increase in born-digital content is being ingested.

Many examples are provided which show the output of specific features of the software in order to display potential areas of benefit for collection institutions. The content provided with said examples have been carefully selected where they do not reveal any private or sensitive information but are still rich with metadata and show the potential of these tools.

## 6.1 Use Case – Digital Forensic Investigation

The first step was selecting the hard drives containing appropriate data. Three hard drives were selected at random from a collection of donations the Computer Archaeology Laboratory at Flinders University has accumulated. Each hard drive was connected via a SATA/IDE to USB adaptor to test the working condition. Once the working drives were discovered, the determination was made on whether they had an adequate level of use and data. This was established by searching the hard drives and looking for indications of high-usage and multiple users. The number of files and directories, including the depth of the directories was a strong indication. If the hard drive contained operating system directories and files, including user directories, this made it a prime candidate offering a rich source of data.

The experiments were conducted on two systems. The first was conducted on a laptop with the following specifications:

Operating system: Windows 10 (version unknown)
CPU: Intel Core i7 – 3630QM, 2.4GHz (up to 3.4GHz) 4-core (8 threads)
Memory: 16 GB DDR3 (1600 MHz)
HDD: 5400 RPM

An upgraded system was utilised to run the experiments again as the original system was performing slow and experiencing lockups and crashes. The new system specifications included:

Operating system: Windows 10 (version unknown)
CPU: Intel i7 8700K – 3.7GHz (4.7GHz Turbo Boost 2.0), 6-core (12 threads)
Cooling: Corsair H100i v2 (Liquid cooled, closed loop)
Memory: 32 GB DDR4 (3000 MHz)
HDD: Seagate FireCuda SSHD (Hybrid, 7200 RPM) / Samsung 960 PRO M.2 SSD

Once each drive was reviewed, exploring the directories, traces of users, and overall data found on the drive, the optimal drive was selected for processing. A RAW (ForensicsWiki, 2017) disk image was created using FTK Imager version 4.2.0 (AccessData, 2017). No writes were permitted on the disk drive and the built in MD5 and SHA-256 checksum generation was utilised. The disk image was reviewed using FTK Imager's viewer, allowing each directory to be selected and expanded, revealing sub-directories and files. The image was also mounted and navigated as a logical drive.

The first investigation involved the use of Bulk_extractor, a tool used for discovering and reporting sensitive data (Garfinkel, 2013). The results are presented in text-based form, made up of text files within a user-defined directory. The output can also be displayed in the Bulk_extractor user interface, known as Bulk_extractor viewer. An additional test was conducted on the results making use of regular expression software to search the output for specific information. This is achieved by using a sequence of characters and symbols that dictate a string or pattern to be searched, where the characters reference what is being searched and the symbols determine how they are interpreted. Bulk_extractor has built in search functionality which can be run through the command-line, or if using the user interface, a regular expression file can be used which will output the search data into a "find.text" file. Although this functionality is built in, in this experiment an additional tool, grepWin, was used which has a more user friendly interface (Küng, 2018).

The second investigation utilises Autopsy, a complete forensic package, used to build a case based on digital evidence to aid in a criminal investigation. Autopsy is the user interface behind The Sleuth Kit (Basis Technology, 2018a). The results are displayed in a range of different formats and visualisations. The multiple viewing options allows data to be seen in its original state as well as text-based, hex, metadata, and other forms of views depending on the type of data. The results from Autopsy serve a purpose of not only the discovery of sensitive data, but also a way to find information more efficiently by ways of categories and sorted user interface elements and visualisation. The contents of an email or document, pictures or video that have been fragmented or deleted, and various information about the system, both hardware and software, in which the hard drive originates from are clearly presented. This includes attached devices, installed software, and other useful information that can be essential for developing emulated environments and determining how the users interacted with said environment.

### 6.1.1 Bulk_extractor and Regular Expression

The Bulk_extractor test was conducted on a hard drive disk image, 13 gigabytes in size that had been used in a family home environment made up of multiple users. The default settings were used to perform the scan. By default, the software utilised maximum performance of the system's CPU. All the scanners were selected except for the "wordlist". An online resource, “Understanding Bulk Extractor Scanners”, which provides descriptions of each scanner is available (Woods, 2018). The wordlist was omitted due to the extensive processing time required as it generates a list of all words that are discovered on the disk image. This is used to discover potential passwords and commonly used words (phrases).

The output for each scanner is presented in a text file and one scanner may be responsible for multiple files. For example, the “**accounts**” scanner generates “**telephone.txt**”, “**ccn.txt**”, “**ccn_track2.txt**”, and “**pii.txt**”. There are two types of output; one contains all the data collected, presented in lines of data containing the targeted item, the other is a histogram that separates the unique values, removing all duplicates and irrelevant surrounding data, leaving only the useful data.

False positives do occur. These are more prominent for certain scanners such as the account scanner where telephone or credit card numbers can be mistaken for other similar values. This is due to similar strings that match the patterns of phone numbers or predicted credit card numbers, but the histograms make it easier to analyse these values and determine the accuracy. Software such as Bulk_extractor can find and sort this information, but the user must still perform their own analysis to carefully validate the accuracy of the returned values. In the event the output is too large for efficient analysis, the user may wish to use additional software of their choosing to allow further processing on the output. This can reduce clutter and extract targeted values into a csv file, for example. Software such as the aforementioned, grepWin, allow the user to efficiently sift through large amounts of data residing in multiple files.

The output is classified under two categories. These are “lines” and “values”. Lines indicates a count of how many lines of data have been generated. These lines may span a few words, or many strings. Each line represents the scanners output, often made up of text, integers, binary, and hexadecimal. This can make it difficult to find the specific item you are after, for example, a sub-webpage within a website URL. The histogram output allows this data to be extracted and displayed.

The “values” category is used for the histograms where each line is a unique value. Unique values may be subsidiary of other values, for example, if a URL search returned ***www.example.com/*** and ***www.example.com/page1***, the second URL is registered as a unique value as are all subsequent pages of ***example.com***. Further example from the Bulk_extractor output of the difference between the lines and values are as follows:

```
http://home.myspace.com/index.cfm?fuseaction=user&MyToken=6f07955c-8712-4f6e-a14c-
6080af7e0a06\x00\x00\x00\x00b7;\x1C\x00\x00\x00\x00\x00\x00\x00\x00http://home.myspa
ce.com/index.cfm?fuseaction=user&MyToken=6f07955c-8712-4f6e-a14c-
6080af7e0a06\x00\x00index[1].cfm\x00\x00

*********************************************************************

http://pav.myspace.com;CBAB9CB80F083E436F6BCC9D9B10396C rofileInfo:
1.0;http://pav.myspace.com;CBAB9CB80F083E436F6BCC9D9B10396C\x0D\x0AContent-
Encoding

*********************************************************************

http://h.live.com/c.gif?RF=&PI=44314&DI=5708&PS=89221 "cleargif"
src="http://h.live.com/c.gif?RF=&PI=44314&DI=5708&PS=89221" width="1" height

*********************************************************************

http://x.myspace.com/images/clear.gif\x00\x00\x00\x00e7\x16N\x00\x00\x00\x00\x0D\xF0\x
AD\x0Bhttp://x.myspace.com/images/clear.gif\x00\xAD\x0Bclear[1].gif\x00

*********************************************************************
```

This displays four **lines** generated by the URL scanner. These are examples of some of the shorter lines. Some of these data may not be interpretable by the user, such as long hexadecimal strings that represent a large number that is typically interpreted by a system, such as web browser. These data may consist of unique strings to be used as login tokens or URL shorteners (redirection keys). Therefore, only some parts of a **line** may be useful to the analyst. This makes the histogram output much more user friendly as the following examples display:

```
n=257  http://www.myspace.com/Modules/Common/Pages/Privacy.aspx

n=251  http://gfx3.mail.live.com/mail/uxp/w2/pr03/HIG/img/h/jewel_24_hover.png

n=249  http://www.myspace.com/Modules/Common/Pages/TermsConditions.aspx

n=247  http://www.myspace.com/Modules/Common/Pages/AboutUs.aspx
```

The “n=” is a count of how many times each item was found in the original output. This means that the first line in the list above was found 257 times. The output will display the results in descending order based on the n count. Therefore, the histograms are much smaller because these additional discoveries are eliminated and reduced to a single value. However, the first line, like many others, still contains multiple entries in the histogram due to values such as:

http://www.myspace.com/Modules/Common/Pages/Privacy.aspx

http://www.myspace.com/Modules/Common/Pages/Privacy.aspx**?MyToken=14fd271e-3f22-4901-810d-c40789821c2e**

The reason for the multiple entries behind this element is due to the nature of the website and how it generates a unique login token (MyToken) every time the user logs in, as seen in **bold**. Many websites are designed differently from one another, therefore the results discovered may vary. A good example of this is how each element on one page of a website was detected as unique values:

http://www.naturalstrategies.com.au/Newsletter.gif

http://www.naturalstrategies.com.au/NewsletterReference.gif

http://www.naturalstrategies.com.au/NewsletterTheme.gif

These detections reveal the different elements the user has explored, allowing the analyst to identify what was viewed and accessed on those websites. As can be seen in the following URL, it reveals the path and the file viewed:

**URL** - http://www.nt.gov.au/pfes/documents/File/police/community/safetyhouse/karama_fun_day_2.jpg

**Path** - /pfes/documents/File/police/community/safetyhouse/

**File** - karama_fun_day_2.jpg

This data may help in establishing information about an individual, their habits, their interests, and correlated with the frequency of their use and other findings, can help in strengthening a profile being developed for the user(s) of the content being analysed.

This process is also viable within a collection institution. If a collection item is centred around an iconic figure, someone of interest to the public, the nature of that collection item may allow for information such as this to be useful. For example, when dealing with poetry or artwork, the artist may have consistent themes throughout their work. Understanding the artists may help identify the potential influence for those themes. The artist’s interests, hobbies, beliefs, traumas, and life experience may all play a part in defining their work.

Another example to consider is based on earlier web browsers when add blockers and pop-op blockers were not as common. Websites could be accidentally loaded by clicking on a hidden window or advertising panel. Even legitimate hyperlinks may contain embedded code to redirect the user to various websites.

With this in mind, if a forensic test was conducted, some websites may be discovered that the user had unintentionally visited. A single occurrence of this discovery may indicate this. Multiple discoveries may not necessarily suggest otherwise if a website they frequently visit constantly forwards them to malicious or questionable websites. However, the use of a website, identified through all the related URLs, can reveal the true nature of their visit. Seeing which elements of the website the user interacted with, including the sub-pages explored, allows this information to be formed.

Human intervention is crucial as patterns can emerge across various Bulk_extractor scanner outputs, requiring analysis, which may be obscure to the untrained eye. Judgements must also be made as accidents do happen regularly, sometimes unavoidable in the case of hidden and embedded malicious code, something that can impact the digital footprint of the user without their knowledge.

Context is important when dealing with user data. What may seem insignificant to one person may be important for another. By capturing a comprehensive set of data, the collection institution ensures it is better prepared for the needs of the public. Context is equally as important in distinguishing the nature of discoveries. For example, some search terms may suggest the need for concern; however, if the user was in a profession, such as psychology or working with troubled children, the search terms may then have the appropriate context. This may include searches for “child abuse” and other similar subjects.

### Scanner results

The disk image processed returned results for majority of the scanners. Out of the total 51 output files, made up of 2 directories (jpeg_carved and unzip_carved) and text files, 38 returned results.

The success of a scanner is determined by how the system was used and what was stored on the hard drive. Different setups may vary in which scanners produce results. In this case, the output files that contained no results were:

- Ccn_track2 + histogram
- Find + histogram
- Gps
- Httplogs
- Ip + histogram
- Kml
- Sqlite_carved
- Unrar_carved
- url_microsoft-live
- vcard

The scanners that were successful returned substantial results and have been split up into multiple values due to an outlier making up a significant majority portion. The total amount of data (lines + values) returned was approximately **8,144,069** lines. This number includes some rounding and false positives, but it is within acceptable parameters. Seven million, three hundred thousand (**7,300,000**) of that number were returned from the “hex” output. The scanner responsible for this output is the “**base16**” scanner.

> *“BASE16 coding, aka hexadecimal or hex code (includes MD5 codes embedded in the data). The primary use of hexadecimal notation is a human-friendly representation of binary-coded values in computing and digital electronics. Hexadecimal is also commonly used to represent computer memory addresses.”* (Woods, 2018)

This leaves a total of **844,069** lines if the hex output is omitted. The results were split into **779,635** lines of data and **64,434** unique values. These results come from a 13-Gigabyte hard drive, last used approximately 13 years ago. Should a larger modern hard drive with active use be processed, these numbers would be significantly larger. This was confirmed in another Bulk_extractor experiment where a single directory of the same size (13GB) was processed. This resulted in different scanners producing results and returned more data overall. This was due to the directory processed being current and the system from which it was derived had been used for many more recent computing tasks. Table 4 displays the original values and Table 5 displays the histogram, which only shows unique values. The results have been split into three brackets: high, medium, and low, indicating the quantity of discoveries. The output files derived from the URL scan whilst not labelled as histograms, contain unique values and therefore have been considered as such.

Table 4 - Bulk_extractor Output (Original)

| Bracket | Scanner output | Lines |
|---|---|---|
| **High** | Hex | 7,300,000 |
| | Domain | 330,000 |
| | URL | 275,000 |
| | Email | 71,000 |
| | Windirs | 50,000 |
| **Medium** | Winpe | 19,000 |
| | Rfc822 | 18,000 |
| | Telephone | 4,000 |
| | Zip | 3,800 |
| | Exif | 3,700 |
| | Winlnk | 2,100 |
| | Json | 2,000 |
| | Unzip_carved | 350 |
| | Winprefetch | 330 |
| **Low** | Ccn | 120 |
| | Ether | 70 |
| | Facebook | 60 |
| | Jpeg_carved | 60 |
| | Rar | 30 |
| | Aes_keys | 10 |
| | Elf | 3 |
| | Pii | 2 |

Table 5 - Bulk_extractor Output (Histogram)

| Bracket | Histograms | Values |
|---|---|---|
| **High** | url_histogram | 46,649 |
| **Medium** | Domain_histogram | 7,540 |
| | url_services | 5,265 |
| | Email_histogram | 2,294 |
| | Telephone_histogram | 933 |
| | Email_domain_histogram | 915 |
| | url_searches | 679 |
| **Low** | Ccn_histogram | 88 |
| | url_facebook-id | 34 |
| | url_facebook-address | 19 |
| | Ether_histogram | 18 |

### Regular expression

Whilst Bulk_extractor sorts its output into named text files and includes the Bulk_extractor viewer, allowing the user to manage the output a little easier, there is still an abundant amount of information to sift through. As has been mentioned, the "regular expression" functionality is built into the software and will output all found instances of the expression into a single text file. This will lengthen processing time considerably as this will need to be processed each time an amendment or change in regular expressions is made.

It is more efficient and convenient to be able to run regular expressions on the output directory, rather than having to run the Bulk_extractor process each time. This allows the alteration of expressions and the ability to add new ones much easier.

The functionality offered with grepWin allows the use of regular expression creation and text-based searching, including the ability to replace found instances with user determined text or strings. There is an output window with two views, "files" and "content". The "files" option will list every file that contains the string or regular expression, how many matches, and the location of the file.

An example of the discoveries is displayed in a popup window when hovering the mouse cursor over the output results. The “content” view displays each line within the searched files, the line it was found on, and showing the discovered string.

An example of a regular expression used for the test is as follows:

([a-zA-Z0-9_\-\.]+)@([a-zA-Z0-9_\-\.]+)\.([a-zA-Z]{2,5})

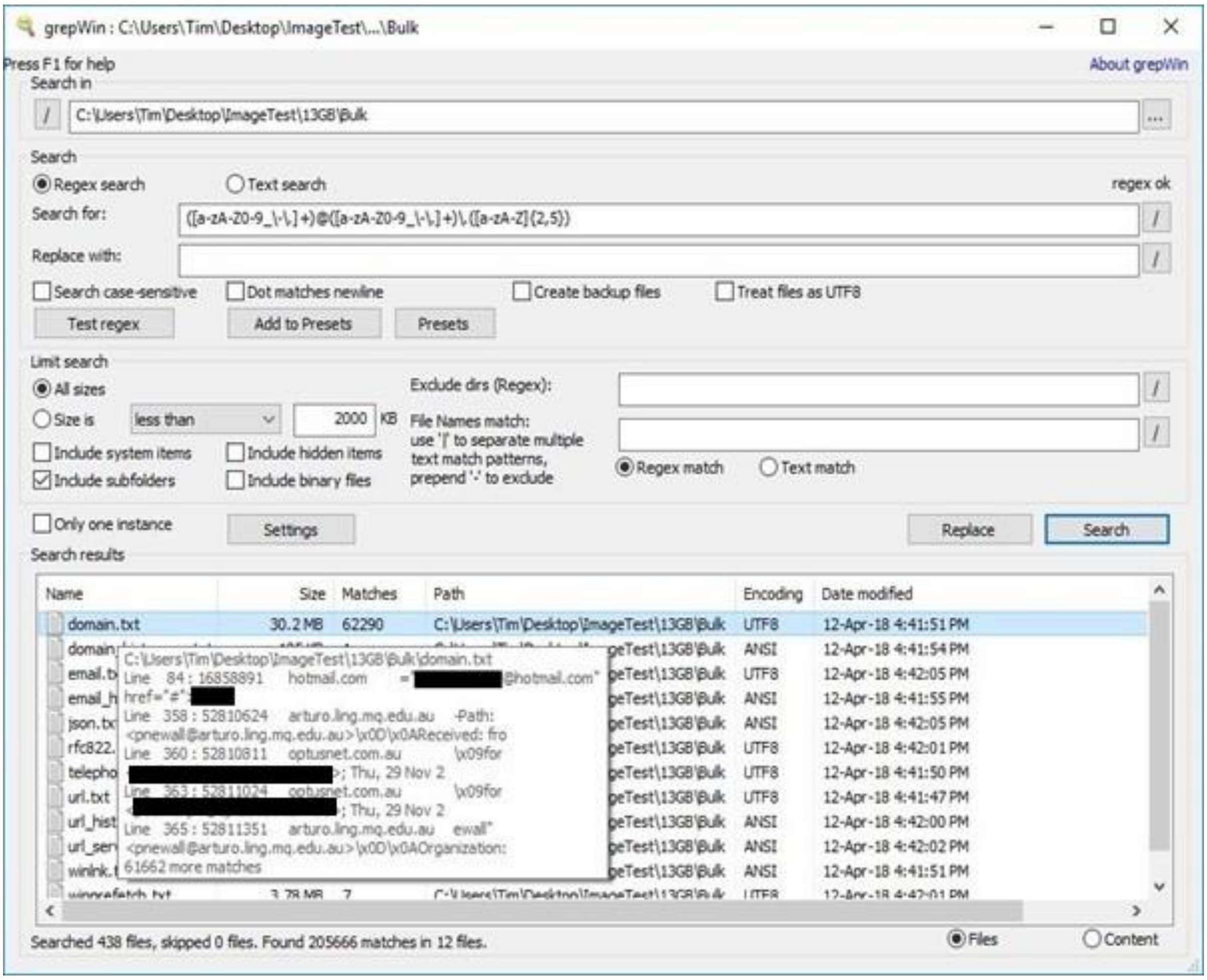
This expression is used to find valid email addresses. The first part indicates characters **A – Z, 0 – 9,** both lower and uppercase are acceptable, including underscores, hyphens, and full stops.

This must then be followed by an **@** symbol and the domain will follow the same parameters as the first half. The last part, which is typically in the form of “.com”, “.org”, “.com.au” may only contain characters **Aa-Zz** in this example, with a minimum of two characters and a maximum of five, indicated by the **{2,5}**.

Some valid emails will fall outside of these parameters and the expression can be tailored accordingly. There are several ways in which an email expression can be written, with different levels of strictness, therefore, many of the expressions that can be created and used look different and may make use of additional notation.

The expression used in this example is missing two common characters that are often found in regular expressions and these are the **^** and **$** symbols. These symbols dictate the start and end of a line; however, they cause issues in this case due to how Bulk_extractor outputs the data. The data are typically surrounded by other strings and are often found in the middle of a line and not always at the beginning. After removing the start and end line symbols, the regular expression performs as intended.  Figure 13 and Figure 14 are displays of the output in both views (with redactions for privacy reasons).

As previously mentioned, regular expression software can be utilised outside of these test parameters. Search and retrieval are fundamental tasks in collection institutions and having the means to automate the process with concise returns makes tools such as grepWin an asset.



Search in: C:\Users\Tim\Desktop\ImageTest\13GB\Bulk

Search
● Regex search ○ Text search regex ok

Search for: ([a-zA-Z0-9_\-\.]+)@([a-zA-Z0-9_\-\.]+)\.([a-zA-Z]{2,5})

Replace with:

□ Search case-sensitive □ Dot matches newline □ Create backup files □ Treat files as UTF8

Limit search
● All sizes
○ Size is less than 2000 KB
□ Include system items □ Include hidden items
☑ Include subfolders □ Include binary files

Exclude dirs (Regex):

File Names match: use '|' to separate multiple text match patterns, prepend '-' to exclude

● Regex match ○ Text match

□ Only one instance

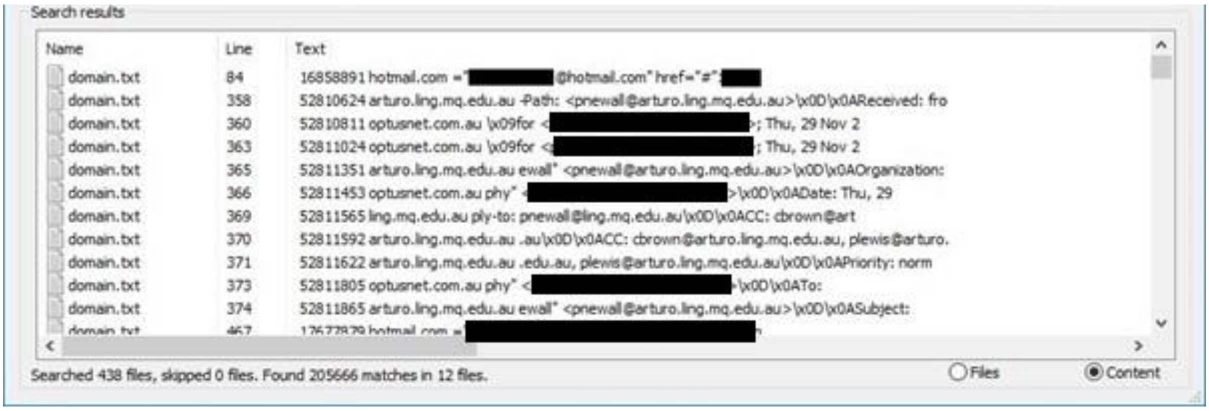
Search results

| Name | Size | Matches | Path | Encoding | Date modified |
|---|---|---|---|---|---|
| domain.txt | 30.2 MB | 62290 | C:\Users\Tim\Desktop\ImageTest\13GB\Bulk | UTF8 | 12-Apr-18 4:41:51 PM |
| domain |  |  | geTest\13GB\Bulk | ANSI | 12-Apr-18 4:41:54 PM |
| email.t |  |  | geTest\13GB\Bulk | UTF8 | 12-Apr-18 4:42:05 PM |
| email_h |  |  | geTest\13GB\Bulk | ANSI | 12-Apr-18 4:41:55 PM |
| json.tx |  |  | geTest\13GB\Bulk | ANSI | 12-Apr-18 4:42:05 PM |
| rfc822. |  |  | geTest\13GB\Bulk | UTF8 | 12-Apr-18 4:42:01 PM |
| telepho |  |  | geTest\13GB\Bulk | UTF8 | 12-Apr-18 4:41:50 PM |
| url.txt |  |  | geTest\13GB\Bulk | UTF8 | 12-Apr-18 4:41:47 PM |
| url_hist |  |  | geTest\13GB\Bulk | ANSI | 12-Apr-18 4:42:00 PM |
| url_ser |  |  | geTest\13GB\Bulk | ANSI | 12-Apr-18 4:42:02 PM |
| winlnk.t |  |  | geTest\13GB\Bulk | ANSI | 12-Apr-18 4:41:51 PM |
| winprefetch.txt | 3.78 MB | 7 | C:\Users\Tim\Desktop\ImageTest\13GB\Bulk | UTF8 | 12-Apr-18 4:42:01 PM |

```
C:\Users\Tim\Desktop\ImageTest\13GB\Bulk\domain.txt
Line   84 : 16858891    hotmail.com     ="[redacted]@hotmail.com"
href="#"[redacted]
Line  358 : 52810624    arturo.ling.mq.edu.au    -Path:
<pnewall@arturo.ling.mq.edu.au>\x0D\x0AReceived: fro
Line  360 : 52810811    optusnet.com.au          \x09for
[redacted]>; Thu, 29 Nov 2
Line  363 : 52811024    optusnet.com.au          \x09for
[redacted]; Thu, 29 Nov 2
Line  365 : 52811351    arturo.ling.mq.edu.au    ewall"
<pnewall@arturo.ling.mq.edu.au>\x0D\x0AOrganization:
61662 more matches
```

Searched 438 files, skipped 0 files. Found 205666 matches in 12 files. ● Files ○ Content

Figure 13 - grepWin, Files View with Pop-Up Display



Search results

| Name | Line | Text |
|---|---|---|
| domain.txt | 84 | 16858891 hotmail.com ="[redacted]@hotmail.com" href="#"[redacted] |
| domain.txt | 358 | 52810624 arturo.ling.mq.edu.au -Path: <pnewall@arturo.ling.mq.edu.au>\x0D\x0AReceived: fro |
| domain.txt | 360 | 52810811 optusnet.com.au \x09for <[redacted]>; Thu, 29 Nov 2 |
| domain.txt | 363 | 52811024 optusnet.com.au \x09for <[redacted]; Thu, 29 Nov 2 |
| domain.txt | 365 | 52811351 arturo.ling.mq.edu.au ewall" <pnewall@arturo.ling.mq.edu.au>\x0D\x0AOrganization: |
| domain.txt | 366 | 52811453 optusnet.com.au phy" <[redacted]>\x0D\x0ADate: Thu, 29 |
| domain.txt | 369 | 52811565 ling.mq.edu.au ply-to: pnewall@ling.mq.edu.au\x0D\x0ACC: cbrown@art |
| domain.txt | 370 | 52811592 arturo.ling.mq.edu.au .au\x0D\x0ACC: cbrown@arturo.ling.mq.edu.au, plewis@arturo. |
| domain.txt | 371 | 52811622 arturo.ling.mq.edu.au .edu.au, plewis@arturo.ling.mq.edu.au\x0D\x0APriority: norm |
| domain.txt | 373 | 52811805 optusnet.com.au phy" <[redacted]>\x0D\x0ATo: |
| domain.txt | 374 | 52811865 arturo.ling.mq.edu.au ewall" <pnewall@arturo.ling.mq.edu.au>\x0D\x0ASubject: |
| domain.txt | 467 | 17677879 hotmail.com =[redacted] |

Searched 438 files, skipped 0 files. Found 205666 matches in 12 files. ○ Files ● Content

Figure 14 - grepWin, Content View

### 6.1.2 Autopsy (The Sleuth Kit)

Autopsy requires significantly more processing time than Bulk_extractor. This puts greater emphasis on the hardware used. The availability of processing power and resources will determine the performance of the task. There are, however, options to select modules individually. This allows the processing to be done in segments as well as the removal of unnecessary processes. The software can still be used whilst processing and the results can be viewed in real-time, but as this an intensive process, performance is significantly impacted. Viewing Autopsy's results during processing may cause lock ups and software crashes. The processing was more efficient when left without any interference.

Due to the visualisations and how the data are displayed in different views, including a timeline and an image gallery, which also includes video, the processing time far exceeds Bulk_extractors text-based functionality, requiring several hours of processing. The visualisations present an informative view on the information gathered.

This section reveals some of the results discovered through the experimentation of selected tools, discussing how they are viable for collection institutions, and showing the benefits of such discoveries.

This experiment undertook two iterations. The second iteration was not performed for the purpose of making a comparison but was done so out of necessity as the old system struggled to run Autopsy, making it difficult to navigate the output and check results.

The data was first processed in Autopsy version 4.6.0, on the initial computer system described previously, and then processed on the second, more modern, computer system. This allowed a comparison to be made on the performance differences. There were complications using Autopsy version 4.9.0 at the time as processing would stop after 10-15 minutes, later to be confirmed as an ingest deadlock. Multiple attempts were made to solve this; however, no solutions were found. A previous version, 4.8.0, was used which worked without issue. Default settings were used in both tests, meaning only two threads were used during ingest. Performance can be increased by setting the number of threads higher, but the maximum recommended is four. This number can be exceeded with more multi-threaded cores, but input/output (I/O) limitations may reduce its effectiveness. Version 4.9.1 was later released, resolving the ingest deadlock issue, but was not re-tested as the output is the primary focus.

The processing time was significantly shorter on the new system as it was much more powerful having 32GB of RAM, opposed to 16GB, and an i7 Intel 6-core processor at 4.7GHz, opposed to an i7 quad-core laptop CPU at 2.4GHz. This still took upwards of 4 hours, but significantly shorter than 7 hours it took on the previous system. The processing times are approximate as the exact processing time is unknown due to interruptions and testing the different modules in stages.

The second iteration of the experiment on the new system utilised more modules and returned more data; therefore, the time difference is not quite accurate, and one can assume should the first test have been identical, the processing time would be longer, making the difference in performance more significant. The difference in versions is a factor to be considered in processing times. This time will also fluctuate depending on system usage at the time of processing. A standalone system with no other processes running, and no interference will result in quicker processing.

Hard drive speeds also have a significant impact on performance. The more modern system utilised a Seagate FireCuda 3.5 inch SSHD, a hybrid solid-state hard drive, 7200 RPM combined with the solid-state technology achieving up to five times the speed of a standard hard drive. The first system utilised an internal notebook hard drive at 5400 RPM. Both systems used the same operating system; Windows 10. Both systems had the latest operating system updates at the time.

Whilst processing times are not directly tied to any research questions, as the goal is to enhance existing preservation workflows with digital forensic software within institutions that are subjected to resource limitations, it is considered relevant. The increase in processing time and the impact on digital preservation workflows is considered and discussed in Chapter 7, Section 7.3 Workflow Enhancements and Visualisations.

### Data Display – Tree Viewer

The first point of focus in Autopsy is the left-side panel where all the data discovered appear and are sorted into types and categories that can be expanded into sub-categories. This is the "Tree Viewer". The data source volumes are found here allowing the user to search through the directories manually as seen in Figure 15.

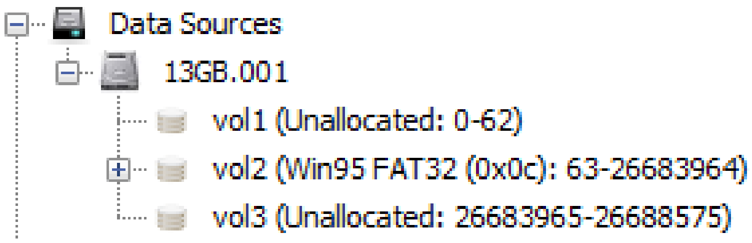
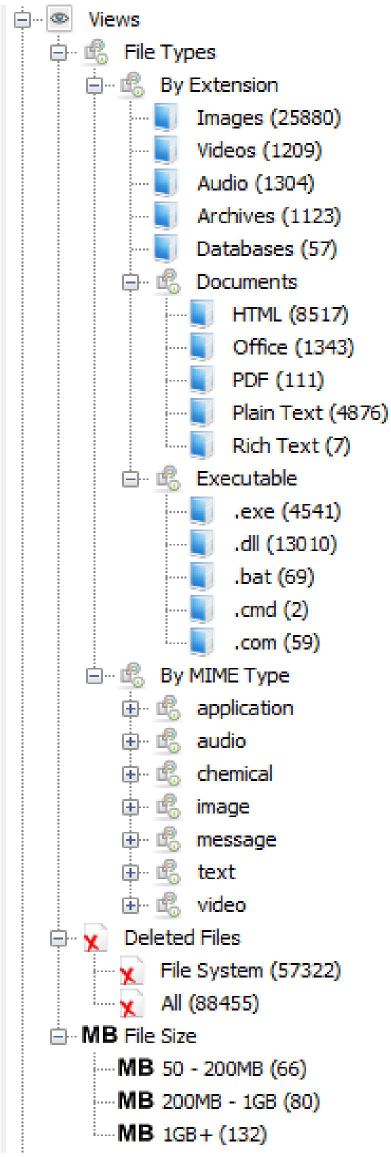

Data Sources
13GB.001
vol1 (Unallocated: 0-62)
vol2 (Win95 FAT32 (0x0c): 63-26683964)
vol3 (Unallocated: 26683965-26688575)

Figure 15 - Autopsy Data Source

Expanding the second volume revealed the directories and their files, including hidden, orphan, deleted files and directories. The remaining categories were sorted by Autopsy and allow for specific searches to be conducted. These fall under three high level categories labelled “views”, “results”, and “tags”. Any reports generated, and items tagged by the user will be sorted under “tags”. The “views” category allows the user to search files based on parameters such as size, type, extension, and deletion status.

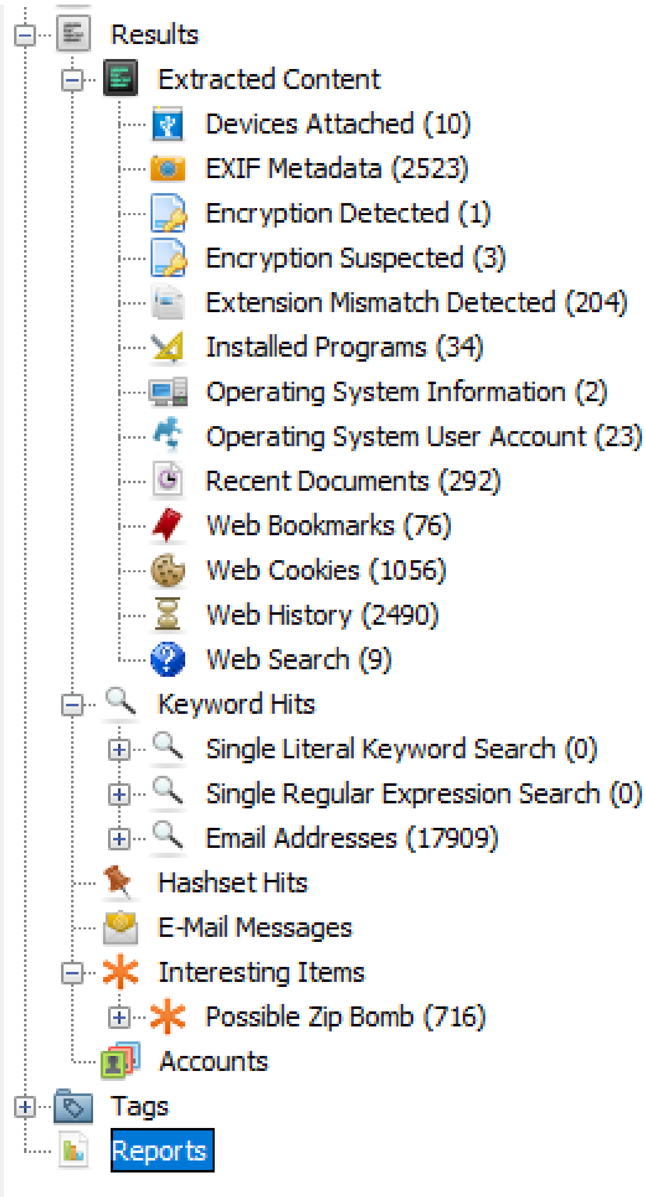


```
Views
  File Types
    By Extension
      Images (25880)
      Videos (1209)
      Audio (1304)
      Archives (1123)
      Databases (57)
      Documents
        HTML (8517)
        Office (1343)
        PDF (111)
        Plain Text (4876)
        Rich Text (7)
      Executable
        .exe (4541)
        .dll (13010)
        .bat (69)
        .cmd (2)
        .com (59)
    By MIME Type
      application
      audio
      chemical
      image
      message
      text
      video
  Deleted Files
    File System (57322)
    All (88455)
  MB File Size
    MB 50 - 200MB (66)
    MB 200MB - 1GB (80)
    MB 1GB+ (132)
```

Figure 16 - Autopsy Views – File Types

Figure 16 displays the different views available which also contain the count of discoveries per category. Viewing a category containing a large number of items increases the chance of performance degradation occurring. For example, trying to view the deleted files, by default, will be limited to ten thousand items and exceeding this will increase load times and reduce stability which can cause the software to lock up. Files are also sorted by MIME types, (Multipurpose Internet Mail Extensions) which are treated as file extensions by web browsers and internet servers (Freed and Kucherawy, 2019).

Figure 17 displays the results and tags which provide information about the system and its users. Here one can establish information about the operating system, each user account, installed programs, and devices that have been attached to that system. This information serves a purpose for both criminal investigation and digital preservation, as information about the original environment can be used when determining the properties of an emulated environment.

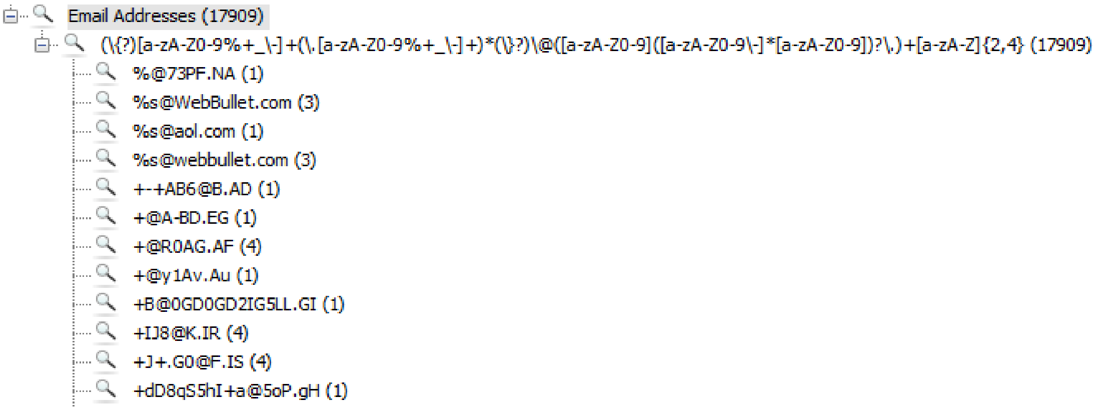


- Results
  - Extracted Content
    - Devices Attached (10)
    - EXIF Metadata (2523)
    - Encryption Detected (1)
    - Encryption Suspected (3)
    - Extension Mismatch Detected (204)
    - Installed Programs (34)
    - Operating System Information (2)
    - Operating System User Account (23)
    - Recent Documents (292)
    - Web Bookmarks (76)
    - Web Cookies (1056)
    - Web History (2490)
    - Web Search (9)
  - Keyword Hits
    - Single Literal Keyword Search (0)
    - Single Regular Expression Search (0)
    - Email Addresses (17909)
  - Hashset Hits
  - E-Mail Messages
  - Interesting Items
    - Possible Zip Bomb (716)
  - Accounts
- Tags
- Reports

Figure 17 - Autopsy Results + Tags

### Keyword and Regular Expression

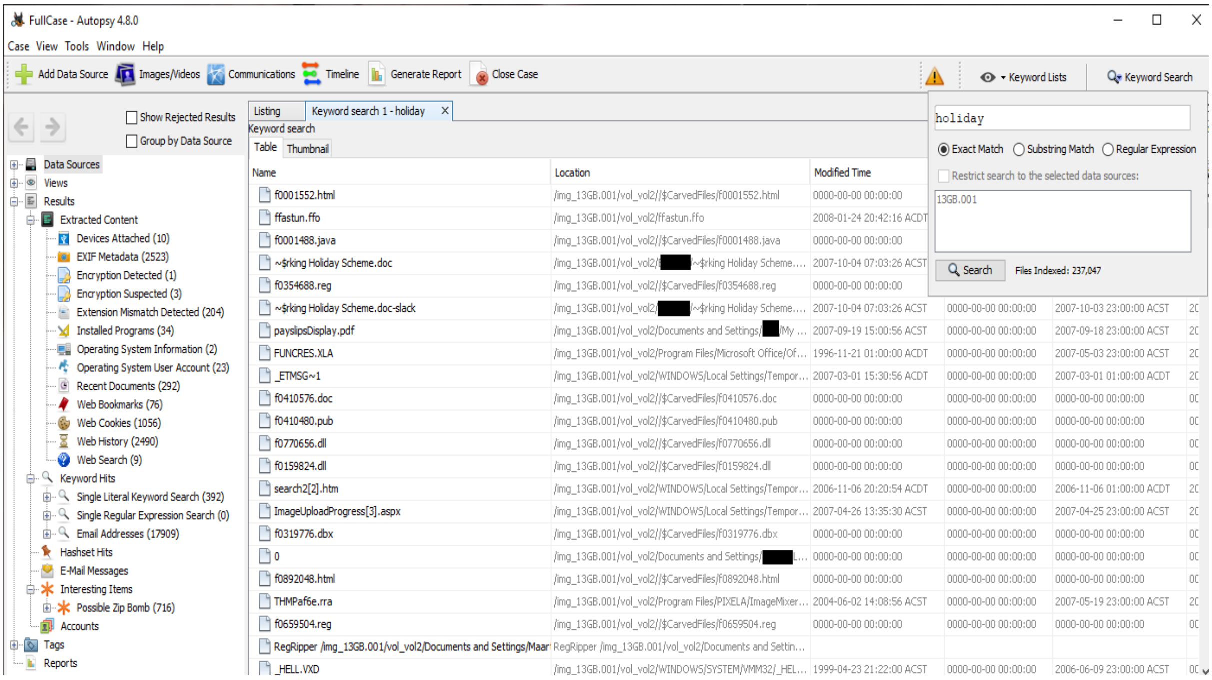
The "Keyword Hits" directory is where keyword and regular expression searches are displayed. By default, a regular expression for email addresses is presented. The regular expression used is quite broad as can be seen in Figure 18 and it will detect many false positives. The extent of this expression may be of interest in some cases, such as criminal investigations where data may intentionally be obscured, but a user-defined expression can be used to reduce the results and increase validity.



```
Email Addresses (17909)
  (\{?)[a-zA-Z0-9%+_\-]+(\.[a-zA-Z0-9%+_\-]+)*(\}?)\@([a-zA-Z0-9]([a-zA-Z0-9\-]*[a-zA-Z0-9])?\.)+[a-zA-Z]{2,4} (17909)
    %@73PF.NA (1)
    %s@WebBullet.com (3)
    %s@aol.com (1)
    %s@webbullet.com (3)
    +-+AB6@B.AD (1)
    +@A-BD.EG (1)
    +@R0AG.AF (4)
    +@y1Av.Au (1)
    +B@0GD0GD2IG5LL.GI (1)
    +IJ8@K.IR (4)
    +J+.G0@F.IS (4)
    +dD8qS5hI+a@5oP.gH (1)
```

Figure 18 - Default Email Regex

Figure 19 displays the keyword search feature where keywords can be entered to define exact matches, substrings, or regular expressions. Some of the results for the keyword “holiday” are displayed.



holiday

● Exact Match ○ Substring Match ○ Regular Expression

□ Restrict search to the selected data sources:

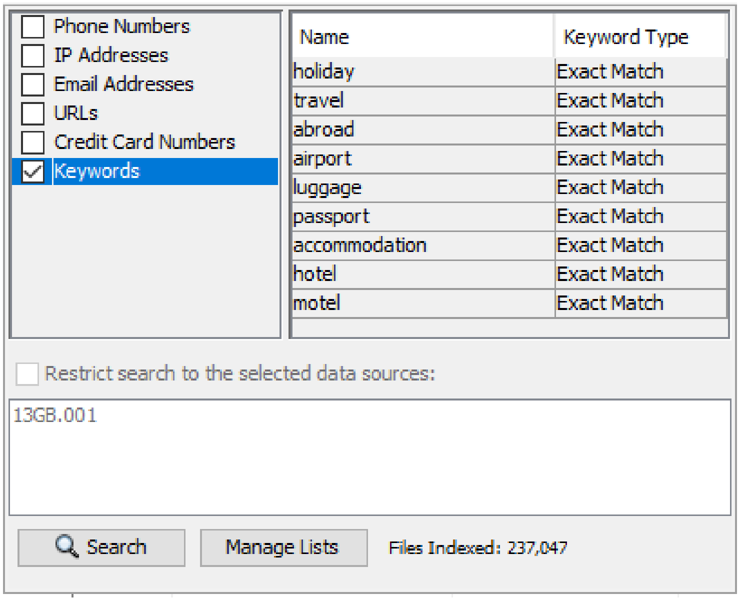
13GB.001

Search Files Indexed: 237,047

| Name | Location | Modified Time | | | |
|---|---|---|---|---|---|
| f0001552.html | /img_13GB.001/vol_vol2//$CarvedFiles/f0001552.html | 0000-00-00 00:00:00 | | | |
| ffastun.ffo | /img_13GB.001/vol_vol2/ffastun.ffo | 2008-01-24 20:42:16 ACDT | | | |
| f0001488.java | /img_13GB.001/vol_vol2//$CarvedFiles/f0001488.java | 0000-00-00 00:00:00 | | | |
| ~$rking Holiday Scheme.doc | /img_13GB.001/vol_vol2/[redacted]/~$rking Holiday Scheme.… | 2007-10-04 07:03:26 ACST | | | |
| f0354688.reg | /img_13GB.001/vol_vol2//$CarvedFiles/f0354688.reg | 0000-00-00 00:00:00 | | | |
| ~$rking Holiday Scheme.doc-slack | /img_13GB.001/vol_vol2/[redacted]/~$rking Holiday Scheme.… | 2007-10-04 07:03:26 ACST | 0000-00-00 00:00:00 | 2007-10-03 23:00:00 ACST | 2C |
| payslipsDisplay.pdf | /img_13GB.001/vol_vol2/Documents and Settings/[redacted]/My … | 2007-09-19 15:00:56 ACST | 0000-00-00 00:00:00 | 2007-09-18 23:00:00 ACST | 2C |
| FUNCRES.XLA | /img_13GB.001/vol_vol2/Program Files/Microsoft Office/Of… | 1996-11-21 01:00:00 ACDT | 0000-00-00 00:00:00 | 2007-05-03 23:00:00 ACST | 2C |
| _ETMSG~1 | /img_13GB.001/vol_vol2/WINDOWS/Local Settings/Tempor… | 2007-03-01 15:30:56 ACDT | 0000-00-00 00:00:00 | 2007-03-01 01:00:00 ACDT | 2C |
| f0410576.doc | /img_13GB.001/vol_vol2//$CarvedFiles/f0410576.doc | 0000-00-00 00:00:00 | 0000-00-00 00:00:00 | 0000-00-00 00:00:00 | 0C |
| f0410480.pub | /img_13GB.001/vol_vol2//$CarvedFiles/f0410480.pub | 0000-00-00 00:00:00 | 0000-00-00 00:00:00 | 0000-00-00 00:00:00 | 0C |
| f0770656.dll | /img_13GB.001/vol_vol2//$CarvedFiles/f0770656.dll | 0000-00-00 00:00:00 | 0000-00-00 00:00:00 | 0000-00-00 00:00:00 | 0C |
| f0159824.dll | /img_13GB.001/vol_vol2//$CarvedFiles/f0159824.dll | 0000-00-00 00:00:00 | 0000-00-00 00:00:00 | 0000-00-00 00:00:00 | 0C |
| search2[2].htm | /img_13GB.001/vol_vol2/WINDOWS/Local Settings/Tempor… | 2006-11-06 20:20:54 ACDT | 0000-00-00 00:00:00 | 2006-11-06 01:00:00 ACDT | 2C |
| ImageUploadProgress[3].aspx | /img_13GB.001/vol_vol2/WINDOWS/Local Settings/Tempor… | 2007-04-26 13:35:30 ACST | 0000-00-00 00:00:00 | 2007-04-25 23:00:00 ACST | 2C |
| f0319776.dbx | /img_13GB.001/vol_vol2//$CarvedFiles/f0319776.dbx | 0000-00-00 00:00:00 | 0000-00-00 00:00:00 | 0000-00-00 00:00:00 | 0C |
| 0 | /img_13GB.001/vol_vol2/Documents and Settings/[redacted]L… | 0000-00-00 00:00:00 | 0000-00-00 00:00:00 | 0000-00-00 00:00:00 | 0C |
| f0892048.html | /img_13GB.001/vol_vol2//$CarvedFiles/f0892048.html | 0000-00-00 00:00:00 | 0000-00-00 00:00:00 | 0000-00-00 00:00:00 | 0C |
| THMPaf6e.rra | /img_13GB.001/vol_vol2/Program Files/PIXELA/ImageMixer… | 2004-06-02 14:08:56 ACST | 0000-00-00 00:00:00 | 2007-05-19 23:00:00 ACST | 2C |
| f0659504.reg | /img_13GB.001/vol_vol2//$CarvedFiles/f0659504.reg | 0000-00-00 00:00:00 | 0000-00-00 00:00:00 | 0000-00-00 00:00:00 | 0C |
| RegRipper /img_13GB.001/vol_vol2/Documents and Settings/Maar | RegRipper /img_13GB.001/vol_vol2/Documents and Settin… | | | | |
| _HELL.VXD | /img_13GB.001/vol_vol2/WINDOWS/SYSTEM/VMM32/_HEL… | 1999-04-23 21:22:00 ACST | 0000-00-00 00:00:00 | 2006-06-09 23:00:00 ACST | 0C |

Figure 19 - Keyword Search

The user may also use a keyword list. Figure 20 displays the default lists which range from phone numbers to credit card numbers. The final selection in the list, “Keywords”, is user generated and contains a list of keywords relating to holiday and travel. Keyword lists can be created, modified, and customised with a mix of exact matches, substrings, and regular expressions.



□ Phone Numbers
□ IP Addresses
□ Email Addresses
□ URLs
□ Credit Card Numbers
☑ Keywords

| Name | Keyword Type |
|---|---|
| holiday | Exact Match |
| travel | Exact Match |
| abroad | Exact Match |
| airport | Exact Match |
| luggage | Exact Match |
| passport | Exact Match |
| accommodation | Exact Match |
| hotel | Exact Match |
| motel | Exact Match |

□ Restrict search to the selected data sources:

13GB.001

Files Indexed: 237,047

Figure 20 - Keyword Lists

### Data Display – Result and Content Viewer

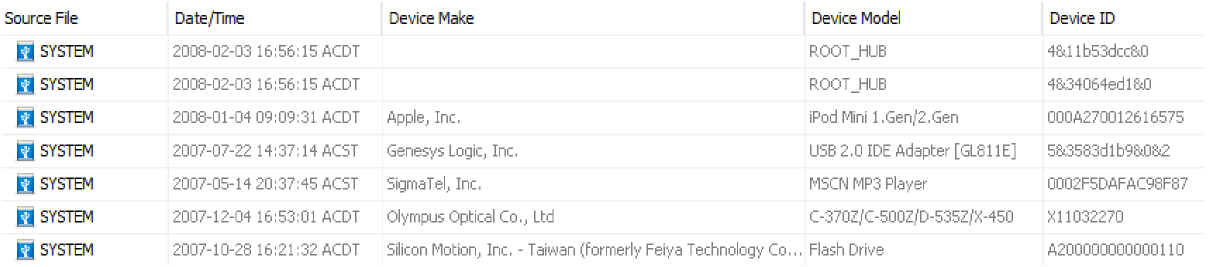
Clicking on any element in the “Tree Viewer” will display the contents in the central panel, known as the “Result Viewer”, allowing the user to investigate each file in detail. Each item selected within this panel will reveal more details in the panel below. This includes information such as the metadata and other specific views depending on the file type. This panel is known as the “Content Viewer” and is where the user can dig deeper into each item discovered on a disk image, giving access to the item’s properties, metadata, string data, hex, and text views. The contents of documents such as Microsoft Office Word, PDF, and Emails can be viewed through these methods, which can be used to establish the workings of an individual, their communications, interests, projects, literature, and other information that can add context to a collection.

The following examples are taken from some of the sub-categories of the results section in the “Tree Viewer”. These examples are from the “Result Viewer” perspective which gives high-level information. The “Content Viewer” will be explored subsequently which gives a lower-level perspective such as metadata, hex, and sorted strings.

The “Devices Attached” displays any discovered devices that have been used on the system. Under this category, the source file, date/time, device make, device model, and device ID are displayed. An example of this in text format is as follows:

- Source File – SYSTEM
- Date/Time – 2008-01-04 09:09:31 ACDT
- Device Make – Apple, Inc.
- Device Model – iPod Mini 1.Gen/2.Gen
- Device ID – 000A230012616575

Figure 21 visually displays the devices discovered. Modern peripherals will show up here such as keyboards, mice, and headsets which now require their own drivers and software which may also include USB ports. This data may allow the user to discover a unique device used on the system which may be required for an executable. For example, if a program was being preserved but was not behaving correctly when trying to use an emulated version, it may be because the incorrect peripheral device is being used. The program may in fact be designed for a unique mouse with features that are not available in modern mice.



| Source File | Date/Time | Device Make | Device Model | Device ID |
|---|---|---|---|---|
| SYSTEM | 2008-02-03 16:56:15 ACDT | | ROOT_HUB | 4&11b53dcc&0 |
| SYSTEM | 2008-02-03 16:56:15 ACDT | | ROOT_HUB | 4&34064ed1&0 |
| SYSTEM | 2008-01-04 09:09:31 ACDT | Apple, Inc. | iPod Mini 1.Gen/2.Gen | 000A270012616575 |
| SYSTEM | 2007-07-22 14:37:14 ACST | Genesys Logic, Inc. | USB 2.0 IDE Adapter [GL811E] | 5&3583d1b9&0&2 |
| SYSTEM | 2007-05-14 20:37:45 ACST | SigmaTel, Inc. | MSCN MP3 Player | 0002F5DAFAC98F87 |
| SYSTEM | 2007-12-04 16:53:01 ACDT | Olympus Optical Co., Ltd | C-370Z/C-500Z/D-535Z/X-450 | X11032270 |
| SYSTEM | 2007-10-28 16:21:32 ACDT | Silicon Motion, Inc. - Taiwan (formerly Feiya Technology Co… | Flash Drive | A200000000000110 |

Figure 21 - Attached Devices

The second category of interest is the “Installed Software”. This category only contains the source file, program name, and the date/time. This information is displayed as follows:

- Source File – SOFTWARE
- Program Name – WebFldrs XP v.9.50.6513
- Date/Time – 2007-10-03 22:44:34 ACST

This is presented in the same format as seen in Figure 21, as is all the information presented in the “Results Viewer”. Information such as this is useful for collection institutions when determining emulation dependencies or determining the software and tools used by an individual to create their work.

The next category, “Operating System Information”, is significant as the operating system is one of the first elements to determine for an emulated environment, including the hardware specifications. Two sources are displayed under the source file column, they are “System” and “Software”.

The system information is displayed as follows: Note ( ) indicates redaction.

- Source File – SYSTEM
- Name – (system name)
- Domain –
- Version – Windows_NT
- Processor Architecture – x86
- Temporary Files Directory - %SystemRoot%\TEMP
- Data Source – (Disk Image)

Software:

- Program Name – Microsoft Windows XP Service Pack 1
- Date/Time – 2007-05-04 10:58:52 ACST
- Path – C:\WINDOWS
- Product ID – 55277-OEM-0049276-21938
- Owner – (User name)

Both sources share the “Tags” field. The processor architecture “x86” refers to 32bit. A 64bit processor will be shown as x64.

The next category relates closely to the operating system information, the “Operating System User Accounts”. Each user (login) is displayed in this category. There are two sources of data. The “Index.dat” contains multiple entries for each user, only displaying the username and a variation shown as “Cookie:(username)”. The other source is labelled as “SOFTWARE”, and this contains the username, the data source, tags, user id, and path.

- Source File – SOFTWARE
- Username – (Username)
- Data Source – (Disk Image)
- Tags –
- User ID – S-1-5-21-1292428093-1283384898-1957994488-1004
- Path - %SystemDrive%\Documents and Settings\(Username)

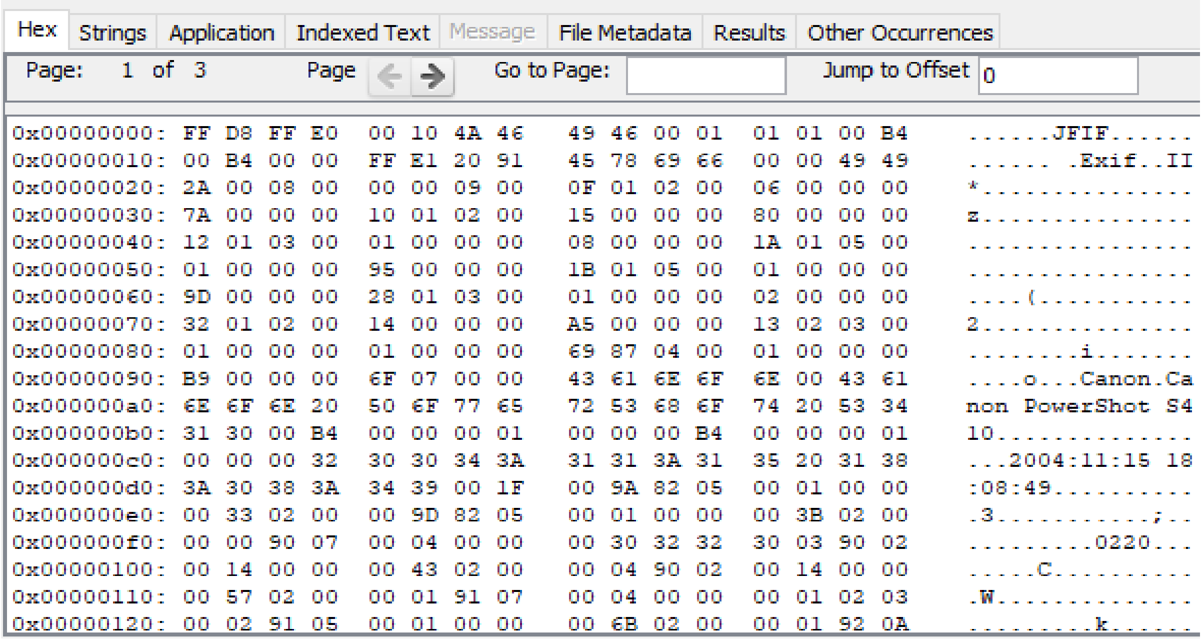
These examples are all extracted from the “Results Viewer” display when selecting an item from the “Tree Viewer”. All examples are real data, hence the need for redaction.

Each of these items is expanded on further in the “Content Viewer”. The data available is determined by the type of file, meaning some of the fields may not be applicable to all file types.

The following examples come from a digital photograph image, taken from the “EXIF Metadata” category in the “Tree Viewer”. The data available in the “Results Viewer” for this item are the source file, date created, device model, device make, data source, size, and the path. The properties of this image are as follows:

- Source File – 250px-Nanzenji_aqueduct_channel[1].jpeg
- Date Created – 2004-11-15 18:08:49 ACDT
- Device Model – Canon Powershot S410
- Device Make – Canon
- Data Source – (Disk Image)
- Size – 34977
- Path - /(disk image)/vol_vol2/Documents and Settings/(Username)/ etc…
- Tags –

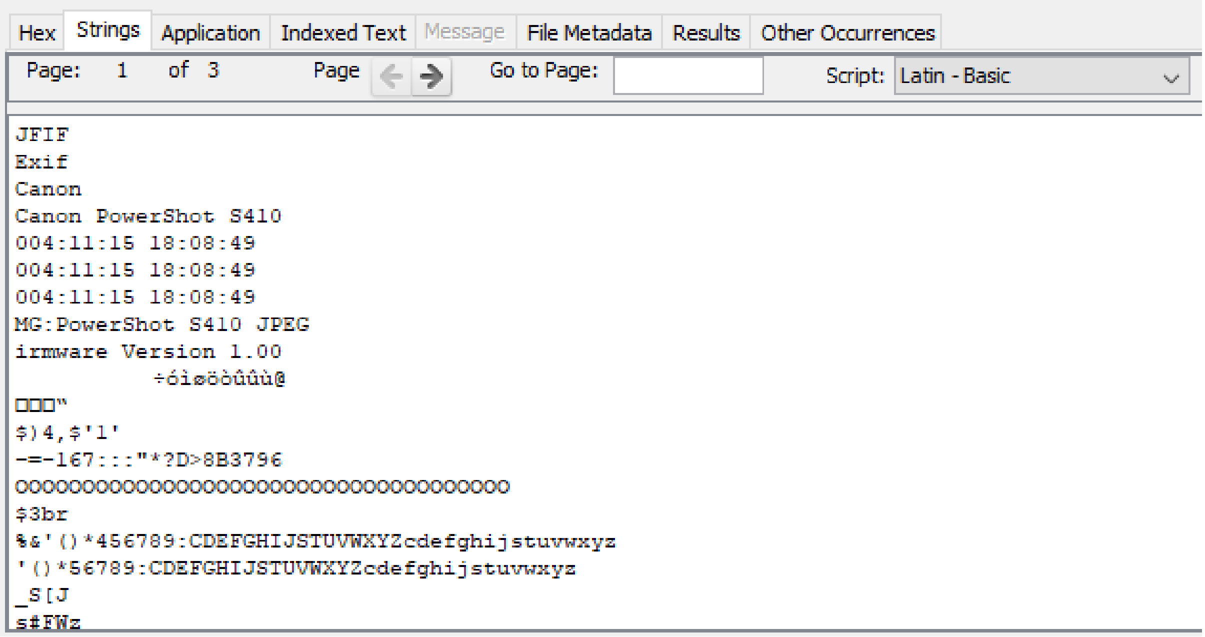
The first tab in the “Content Viewer” (Figure 22) is the “Hex” tab. This presents the user with a hex view, displaying the binary data, represented in hexadecimal, giving a raw data perspective of selected files.



```
0x00000000: FF D8 FF E0  00 10 4A 46  49 46 00 01  01 01 00 B4   ......JFIF......
0x00000010: 00 B4 00 00  FF E1 20 91  45 78 69 66  00 00 49 49   ...... .Exif..II
0x00000020: 2A 00 08 00  00 00 09 00  0F 01 02 00  06 00 00 00   *...............
0x00000030: 7A 00 00 00  10 01 02 00  15 00 00 00  80 00 00 00   z...............
0x00000040: 12 01 03 00  01 00 00 00  08 00 00 00  1A 01 05 00   ................
0x00000050: 01 00 00 00  95 00 00 00  1B 01 05 00  01 00 00 00   ................
0x00000060: 9D 00 00 00  28 01 03 00  01 00 00 00  02 00 00 00   ....(...........
0x00000070: 32 01 02 00  14 00 00 00  A5 00 00 00  13 02 03 00   2...............
0x00000080: 01 00 00 00  01 00 00 00  69 87 04 00  01 00 00 00   ........i.......
0x00000090: B9 00 00 00  6F 07 00 00  43 61 6E 6F  6E 00 43 61   ....o...Canon.Ca
0x000000a0: 6E 6F 6E 20  50 6F 77 65  72 53 68 6F  74 20 53 34   non PowerShot S4
0x000000b0: 31 30 00 B4  00 00 00 01  00 00 00 B4  00 00 00 01   10..............
0x000000c0: 00 00 00 32  30 30 34 3A  31 31 3A 31  35 20 31 38   ...2004:11:15 18
0x000000d0: 3A 30 38 3A  34 39 00 1F  00 9A 82 05  00 01 00 00   :08:49..........
0x000000e0: 00 33 02 00  00 9D 82 05  00 01 00 00  00 3B 02 00   .3...........;..
0x000000f0: 00 00 90 07  00 04 00 00  00 30 32 32  30 03 90 02   .........0220...
0x00000100: 00 14 00 00  00 43 02 00  00 04 90 02  00 14 00 00   .....C..........
0x00000110: 00 57 02 00  00 01 91 07  00 04 00 00  00 01 02 03   .W..............
0x00000120: 00 02 91 05  00 01 00 00  00 6B 02 00  00 01 92 0A   .........k......
```

Figure 22 - Content Viewer - Hex Tab

The “Strings” tab (Figure 23) displays all the text-based strings with the option to change the script which allows non-Latin alphabets.



Page: 1 of 3

Script: Latin - Basic

```
JFIF
Exif
Canon
Canon PowerShot S410
004:11:15 18:08:49
004:11:15 18:08:49
004:11:15 18:08:49
MG:PowerShot S410 JPEG
irmware Version 1.00
          ÷óìøöòûûù@
□□□"
$)4,$'1'
-=-167:::"*?D>8B3796
0000000000000000000000000000000000000
$3br
%&'()*456789:CDEFGHIJSTUVWXYZcdefghijstuvwxyz
'()*56789:CDEFGHIJSTUVWXYZcdefghijstuvwxyz
_S[J
s#FWz
```

Figure 23 - Content Viewer - Strings Tab

The “Application” tab (Figure 24) visually displays files. Most image files can be displayed, even fragmented and partial images will reveal some of the original content. SQLite database tables can be accessed and “plist” file data can also be viewed. A “plist” file is used by macOS applications containing properties and configurations (Fileinfo, 2017).

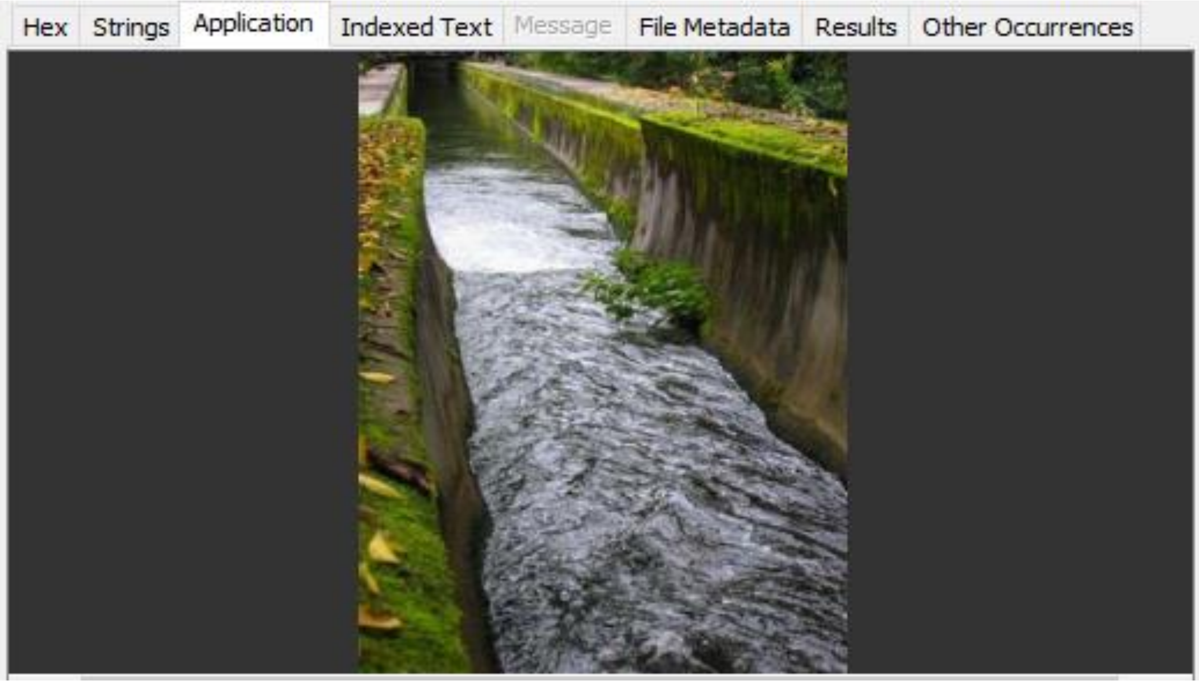
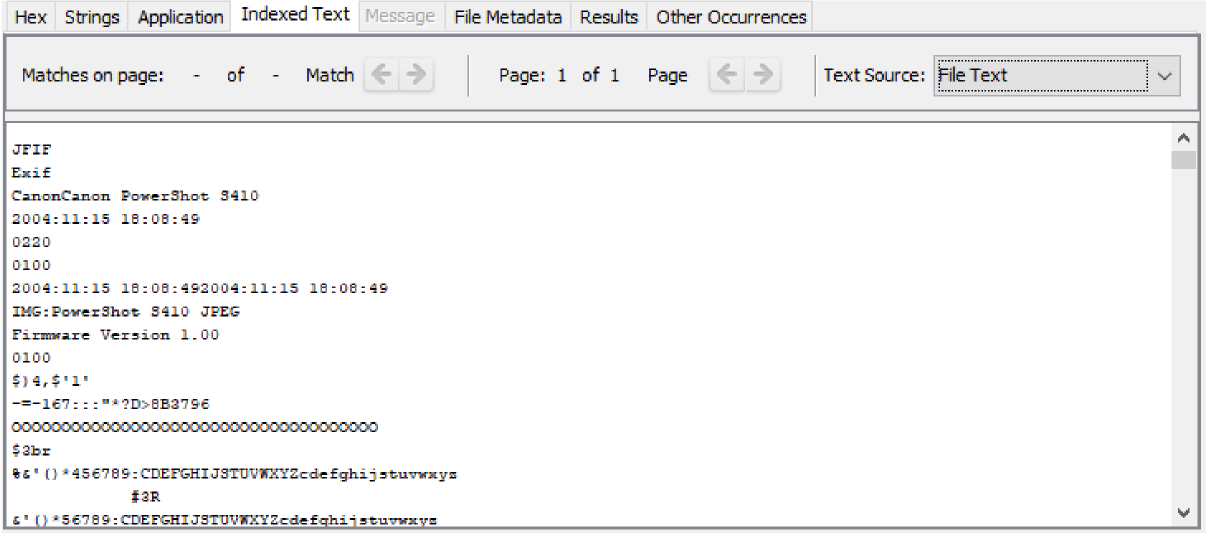


Figure 24 - Content Viewer – Application

Examples for the SQLite and the plist data can be found on the Autopsy user documentation web page for the “Content Viewer” (Basis Technology, 2018b). Other examples for all the tabs in the “Content Viewer” can also be found within the user documentation, but for consistency, all outputs will be shown based on the test data where data are present.

The “Indexed Text” (Figure 25) display is similar to the “Strings” tab shown in Figure 23, the results are almost identical. However, the results from this tab have been indexed by the keyword module. By switching the text source from “file text” to “result text” in the drop-down menu, the results are matched based on the source. For example, by switching to the “result text” for the selected file, the matched results are simplified and return:

Date Created : 2004-11-15 18:08:49 ACDT
Device Model : Canon PowerShot S410
Device Make : Canon



```
JFIF
Exif
CanonCanon PowerShot S410
2004:11:15 18:08:49
0220
0100
2004:11:15 18:08:492004:11:15 18:08:49
IMG:PowerShot S410 JPEG
Firmware Version 1.00
0100
$)4,$'1'
-=-167:::"*?D>8B3796
OOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOO
$3br
%&'()*456789:CDEFGHIJSTUVWXYZcdefghijstuvwxyz
            $3R
&'()*56789:CDEFGHIJSTUVWXYZcdefghijstuvwxyz
```

Figure 25 - Content Viewer - Indexed Text

The “Message” tab is not available for the selected item (jpeg image). This tab shows the contents and details for emails and mobile phone messages (SMS). Examples are available in the online user documentation.

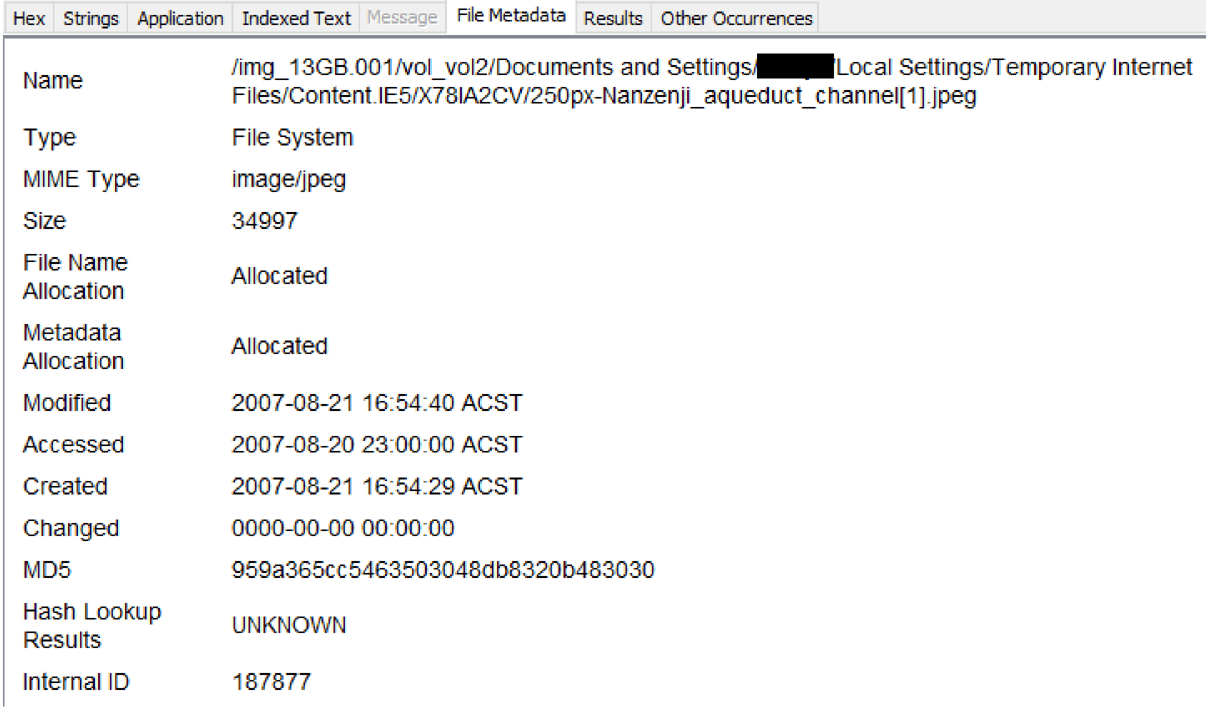
The “File Metadata” tab is made up of two components. The first is the basic file metadata as seen in Figure 26 and the second is the output from the “istat” which is part of The Sleuth Kit. The istat output displays statistics and details about the metadata structure as follows:

```
Directory Entry: 185446629
Allocated
File Attributes: File, Archive
```

Size: 34997
Name: 250PX-~2.JPE

Directory Entry Times:
Written: 2007-08-21 17:54:40 (Cen. Australia Standard Time)
Accessed: 2007-08-21 00:00:00 (Cen. Australia Standard Time)
Created: 2007-08-21 17:54:29 (Cen. Australia Standard Time)

Sectors:
Staring address: 5591294, length: 69

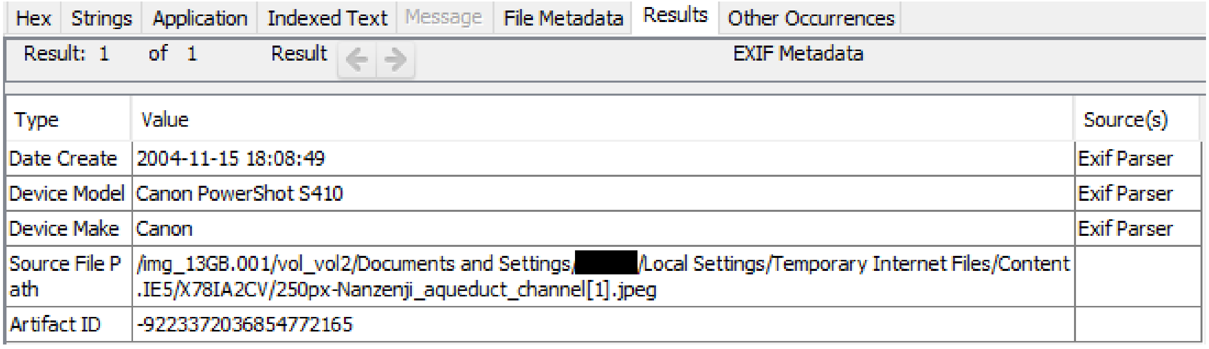


| | |
|---|---|
| Name | /img_13GB.001/vol_vol2/Documents and Settings/[redacted]/Local Settings/Temporary Internet Files/Content.IE5/X78IA2CV/250px-Nanzenji_aqueduct_channel[1].jpeg |
| Type | File System |
| MIME Type | image/jpeg |
| Size | 34997 |
| File Name Allocation | Allocated |
| Metadata Allocation | Allocated |
| Modified | 2007-08-21 16:54:40 ACST |
| Accessed | 2007-08-20 23:00:00 ACST |
| Created | 2007-08-21 16:54:29 ACST |
| Changed | 0000-00-00 00:00:00 |
| MD5 | 959a365cc5463503048db8320b483030 |
| Hash Lookup Results | UNKNOWN |
| Internal ID | 187877 |

**Figure 26 - Content Viewer - File Metadata**

The “Results” tab (Figure 27) is dynamic and changes depending on the type of file selected from the “Results Tree”. The source(s) are based on the module used to generate the data on the file selected. For example, as the selected item is from the “Exif Metadata” category in the “Results Tree”, the “Exif Parser” is the module responsible and is therefore the source. Should a webpage bookmark be selected from the “Web Bookmarks” category, the source will change to “RecentActivity”.

The “Other Occurrences” tab was not applicable for this data, but it is used to reveal where the selected file or results have occurred in other places and can be used to correlate data. The “Central Repository” adds additional functionality to this tab if enabled. Refer to the Autopsy user guide (Basis Technology, 2018c).



| Type | Value | Source(s) |
|---|---|---|
| Date Create | 2004-11-15 18:08:49 | Exif Parser |
| Device Model | Canon PowerShot S410 | Exif Parser |
| Device Make | Canon | Exif Parser |
| Source File Path | /img_13GB.001/vol_vol2/Documents and Settings/[redacted]/Local Settings/Temporary Internet Files/Content.IE5/X78IA2CV/250px-Nanzenji_aqueduct_channel[1].jpeg | |
| Artifact ID | -9223372036854772165 | |

Figure 27 - Content Viewer – Results

### Visualisation

Autopsy offers three distinct forms of visualisation. The first is a gallery that allows the user to navigate images and videos found on the disk image. The gallery offers multiple ways to sort and view images. This is primarily used for forensic analysts to review and tag images based on the five categories of illicit material, displayed as follows:

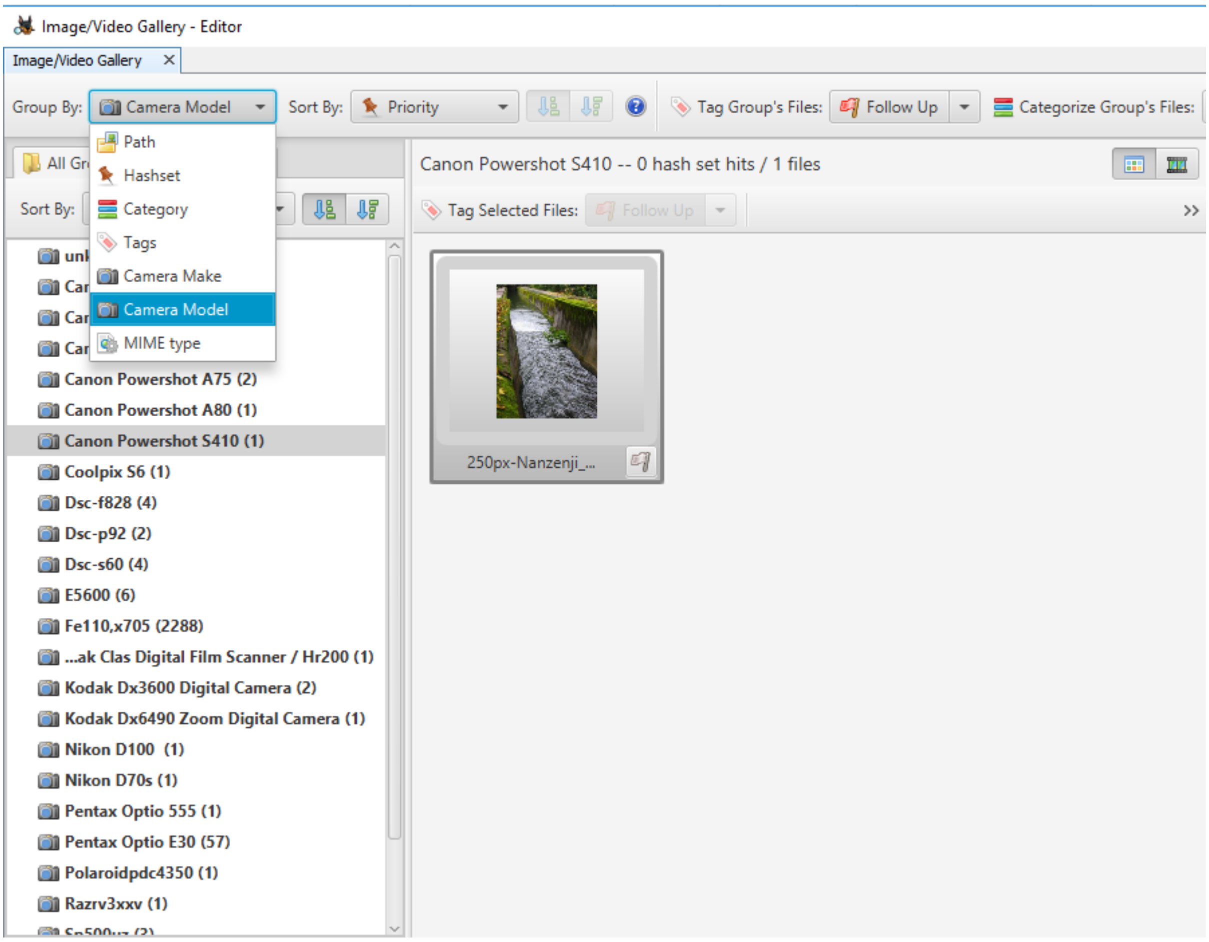
- CAT-1: Child Exploitation (Illegal)
- CAT-2: Child Exploitation (Non-Illegal/Age Difficult)
- CAT-3: CGI/Animation (Child Exploitive)
- CAT-4: Exemplar/Comparison (Internal Use Only)
- CAT-5: Non-pertinent
- CAT-0: Uncategorized

It can also be used for other purposes such as being able to quickly navigate through all the images on a disk image and sort them by groups. Figure 28 illustrates how images and videos may be selected via different groups. This example is based on camera models, which can also be refined into camera make, for a more refined list.  Having this capability is useful for cataloguing digital images by having them sorted by the camera responsible for each photo. This allows an additional level of sorting and organisation of files during the preservation process.

The default view is the “Path” view which displays a directory tree. All images found on the disk image can be accessed through the original, sorted directories, with tagging and exporting capabilities. The user may also choose to go to the source of an image, which will direct the user to its location within the “Results Viewer”.

The second visualisation is “Communications”. This gives an overview of the communications for a case, including the most frequent accounts, communications between accounts, time-stamps, and it allows these data to be browsed as a list or visualisation map as seen in Figure 29 and Figure 30. Through this visualisation, one can navigate the accounts and the communications sent between internal and external users to establish relationships which may lead to new pathways to investigate. This is an important process when cataloguing an iconic or public figure, should the user be ingesting material from a personal device that may have been discovered and donated to the institution. The correspondence to and from the person of interest may reveal data that changes the nature of the collection which is something that must be addressed for completeness and accuracy.

No results were generated for the disk image used. Figure 29 and Figure 30 from the Autopsy communication and visualisation documentation illustrate the "Communications" interface. (Basis Technology, 2018d)



Image/Video Gallery - Editor

Group By: Camera Model

Sort By: Priority

Tag Group's Files: Follow Up

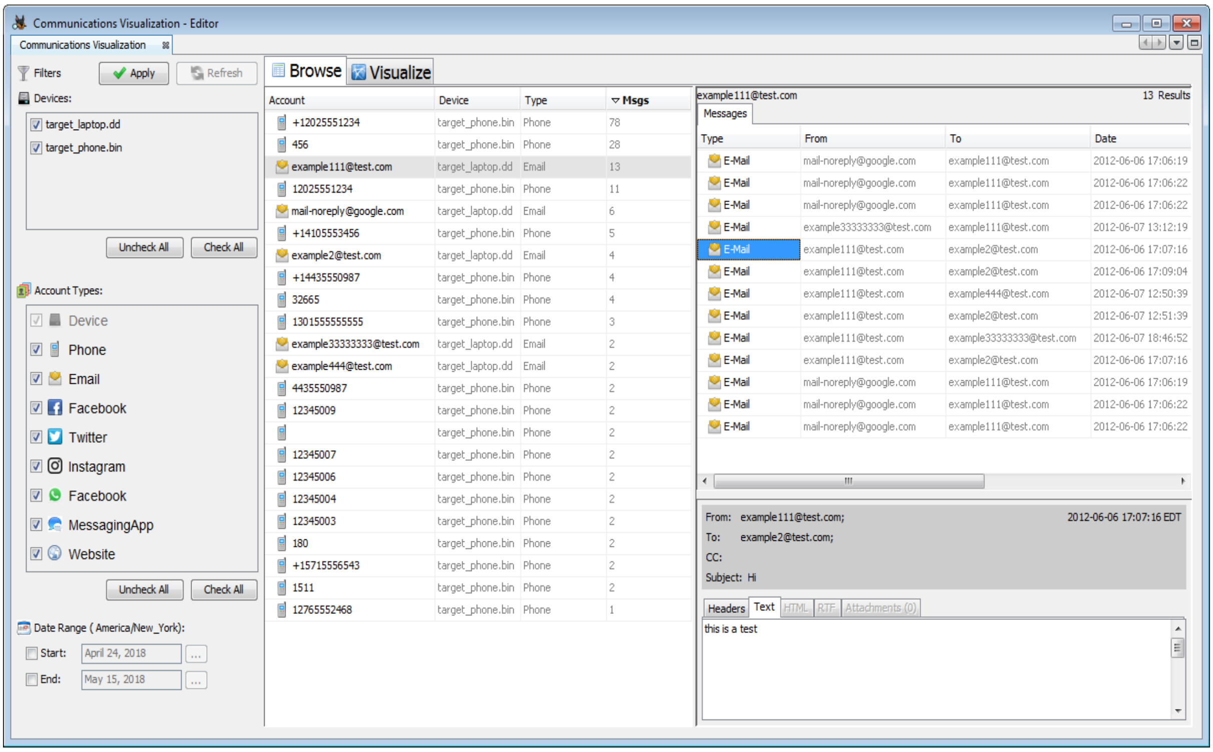
Canon Powershot S410 -- 0 hash set hits / 1 files

Tag Selected Files: Follow Up

- Canon Powershot A75 (2)
- Canon Powershot A80 (1)
- Canon Powershot S410 (1)
- Coolpix S6 (1)
- Dsc-f828 (4)
- Dsc-p92 (2)
- Dsc-s60 (4)
- E5600 (6)
- Fe110,x705 (2288)
- ...ak Clas Digital Film Scanner / Hr200 (1)
- Kodak Dx3600 Digital Camera (2)
- Kodak Dx6490 Zoom Digital Camera (1)
- Nikon D100 (1)
- Nikon D70s (1)
- Pentax Optio 555 (1)
- Pentax Optio E30 (57)
- Polaroidpdc4350 (1)
- Razrv3xxv (1)

250px-Nanzenji_...

Figure 28 - Gallery Groups



Devices:

☑ target_laptop.dd
☑ target_phone.bin

Account Types:

☑ Device
☑ Phone
☑ Email
☑ Facebook
☑ Twitter
☑ Instagram
☑ Facebook
☑ MessagingApp
☑ Website

Date Range ( America/New_York):

☐ Start: April 24, 2018
☐ End: May 15, 2018

| Account | Device | Type | ▽ Msgs |
|---|---|---|---|
| +12025551234 | target_phone.bin | Phone | 78 |
| 456 | target_phone.bin | Phone | 28 |
| example111@test.com | target_laptop.dd | Email | 13 |
| 12025551234 | target_phone.bin | Phone | 11 |
| mail-noreply@google.com | target_laptop.dd | Email | 6 |
| +14105553456 | target_phone.bin | Phone | 5 |
| example2@test.com | target_laptop.dd | Email | 4 |
| +14435550987 | target_phone.bin | Phone | 4 |
| 32665 | target_phone.bin | Phone | 4 |
| 1301555555555 | target_phone.bin | Phone | 3 |
| example33333333@test.com | target_laptop.dd | Email | 2 |
| example444@test.com | target_laptop.dd | Email | 2 |
| 4435550987 | target_phone.bin | Phone | 2 |
| 12345009 | target_phone.bin | Phone | 2 |
|  | target_phone.bin | Phone | 2 |
| 12345007 | target_phone.bin | Phone | 2 |
| 12345006 | target_phone.bin | Phone | 2 |
| 12345004 | target_phone.bin | Phone | 2 |
| 12345003 | target_phone.bin | Phone | 2 |
| 180 | target_phone.bin | Phone | 2 |
| +15715556543 | target_phone.bin | Phone | 2 |
| 1511 | target_phone.bin | Phone | 2 |
| 12765552468 | target_phone.bin | Phone | 1 |

example111@test.com 13 Results

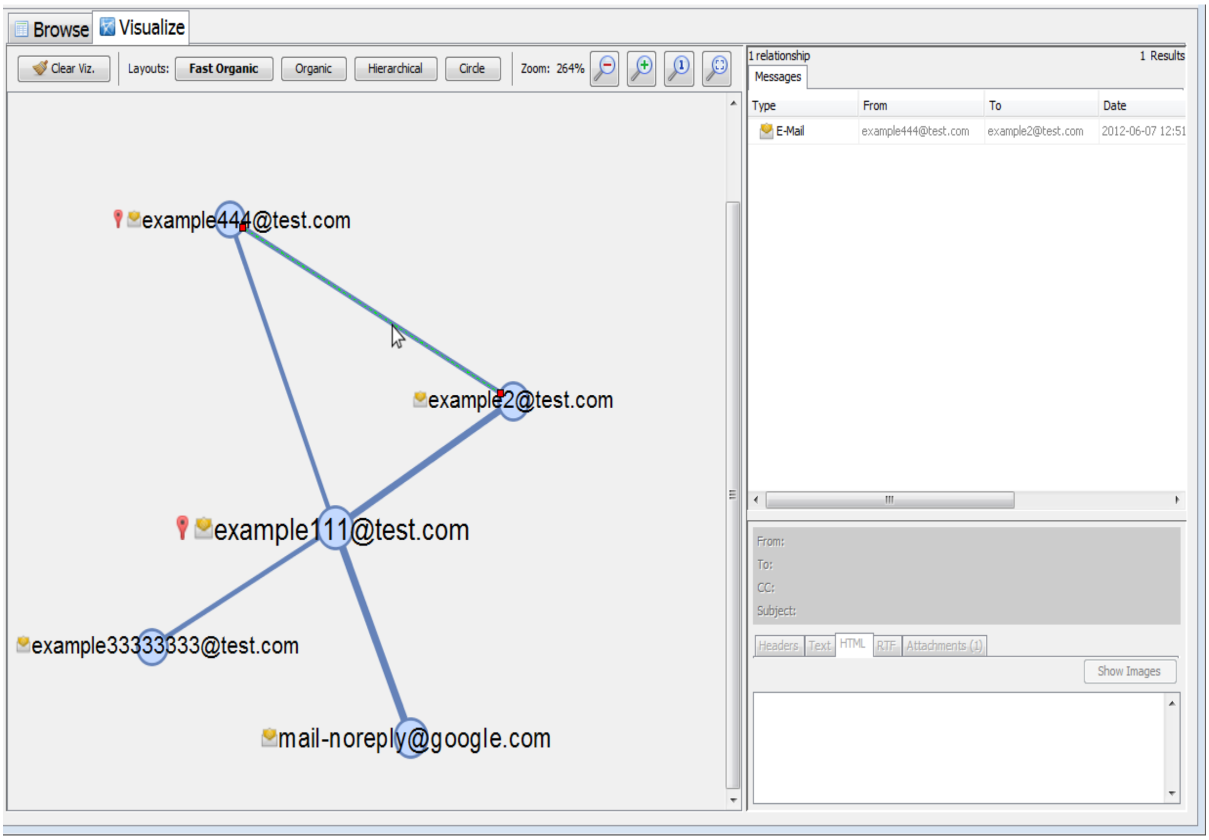
Messages

| Type | From | To | Date |
|---|---|---|---|
| E-Mail | mail-noreply@google.com | example111@test.com | 2012-06-06 17:06:19 |
| E-Mail | mail-noreply@google.com | example111@test.com | 2012-06-06 17:06:22 |
| E-Mail | mail-noreply@google.com | example111@test.com | 2012-06-06 17:06:22 |
| E-Mail | example33333333@test.com | example111@test.com | 2012-06-07 13:12:19 |
| E-Mail | example111@test.com | example2@test.com | 2012-06-06 17:07:16 |
| E-Mail | example111@test.com | example2@test.com | 2012-06-06 17:09:04 |
| E-Mail | example111@test.com | example444@test.com | 2012-06-07 12:50:39 |
| E-Mail | example111@test.com | example2@test.com | 2012-06-07 12:51:39 |
| E-Mail | example111@test.com | example33333333@test.com | 2012-06-07 18:46:52 |
| E-Mail | example111@test.com | example2@test.com | 2012-06-06 17:07:16 |
| E-Mail | mail-noreply@google.com | example111@test.com | 2012-06-06 17:06:19 |
| E-Mail | mail-noreply@google.com | example111@test.com | 2012-06-06 17:06:22 |
| E-Mail | mail-noreply@google.com | example111@test.com | 2012-06-06 17:06:22 |

From: example111@test.com; 2012-06-06 17:07:16 EDT
To: example2@test.com;
CC:
Subject: Hi

this is a test

Figure 29 - Communications List View (Basis Technology, 2018d)



1 relationship 1 Results

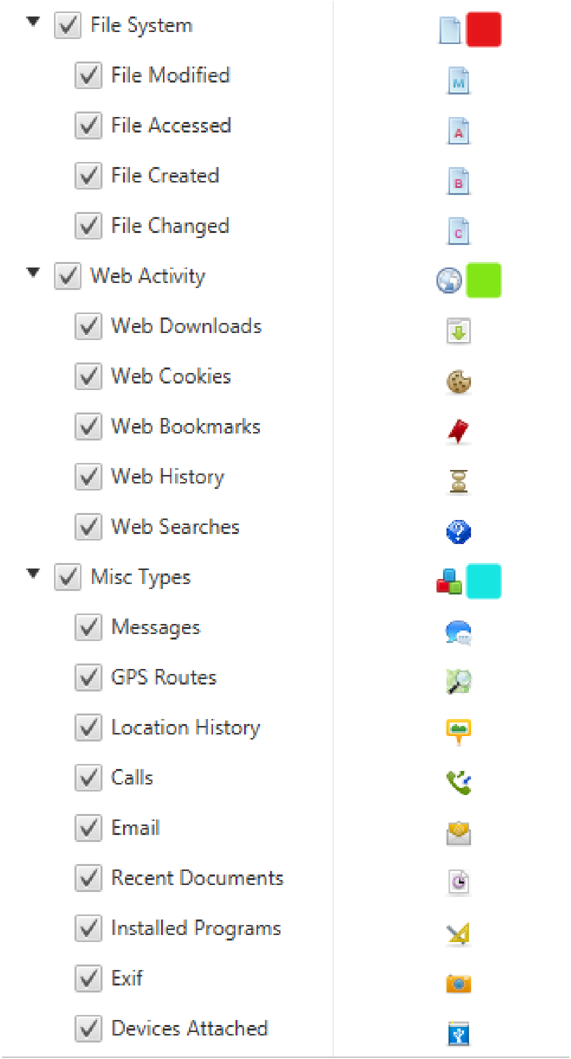
Messages

| Type | From | To | Date |
|---|---|---|---|
| E-Mail | example444@test.com | example2@test.com | 2012-06-07 12:51 |

From:
To:
CC:
Subject:

Figure 30 - Communications Visualised (Basis Technology, 2018d)

The third visualisation is a timeline that offers time-based visualisations. Through multiple views, this feature allows the user to navigate the timeline based on events or activities that surround files and when they occurred using a colour coded and icon based key as seen in Figure 31:



- ☑ File System 🟥
  - ☑ File Modified
  - ☑ File Accessed
  - ☑ File Created
  - ☑ File Changed
- ☑ Web Activity 🟩
  - ☑ Web Downloads
  - ☑ Web Cookies
  - ☑ Web Bookmarks
  - ☑ Web History
  - ☑ Web Searches
- ☑ Misc Types 🟦
  - ☑ Messages
  - ☑ GPS Routes
  - ☑ Location History
  - ☑ Calls
  - ☑ Email
  - ☑ Recent Documents
  - ☑ Installed Programs
  - ☑ Exif
  - ☑ Devices Attached

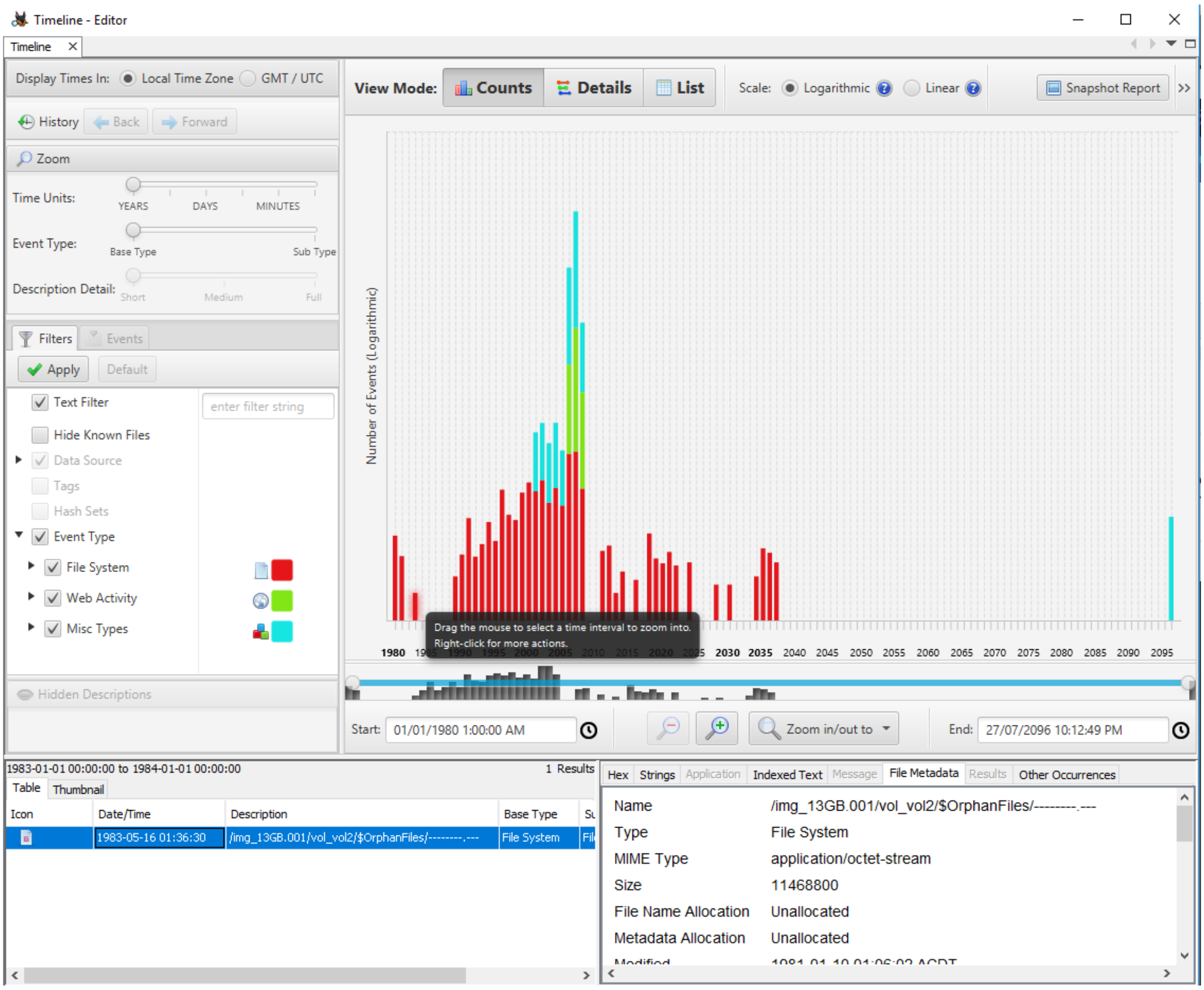
**Figure 31 - Timeline Event Legend**

The use case details presented in the Autopsy timeline documentation (Basis Technology, 2018e) offer a forensic perspective of how the timeline feature was designed, such as:

- "When did major web activity occur on a system?
- When were external devices plugged into the system?
- When were pictures with EXIF information added?
- What websites were accessed that resulted in file system modifications immediately after?"

These concepts can easily be re-interpreted through a collection institution perspective. Determining provenance and the history of a file, including the change history and custodianship, are some of main objectives of digital preservation. Having this timeline can help in determining such information, especially across a multitude of files where events may have occurred at different times. Seeing other events that occurred around the same time as the event in question can lead to further discovery and may reveal correlations that were not obvious prior to this visualisation.

There are three views presented in the timeline. The first view is the “Counts” view, where the colour-coded bar chart is displayed against a timeline and displays each event type, the colour, and the associated icon. Each colour represents a different type of activity.



Timeline - Editor

Display Times In: ● Local Time Zone ○ GMT / UTC

View Mode: Counts | Details | List

Scale: ● Logarithmic ○ Linear

Time Units: YEARS

Event Type: Base Type

Description Detail: Short

☑ Text Filter — enter filter string
☐ Hide Known Files
☑ Data Source
☐ Tags
☐ Hash Sets
☑ Event Type
☑ File System 🟥
☑ Web Activity 🟩
☑ Misc Types 🟦

Drag the mouse to select a time interval to zoom into.
Right-click for more actions.

Start: 01/01/1980 1:00:00 AM

End: 27/07/2096 10:12:49 PM

1983-01-01 00:00:00 to 1984-01-01 00:00:00 — 1 Results

| Icon | Date/Time | Description | Base Type | Su |
|---|---|---|---|---|
|  | 1983-05-16 01:36:30 | /img_13GB.001/vol_vol2/$OrphanFiles/--------.--- | File System | Fil |

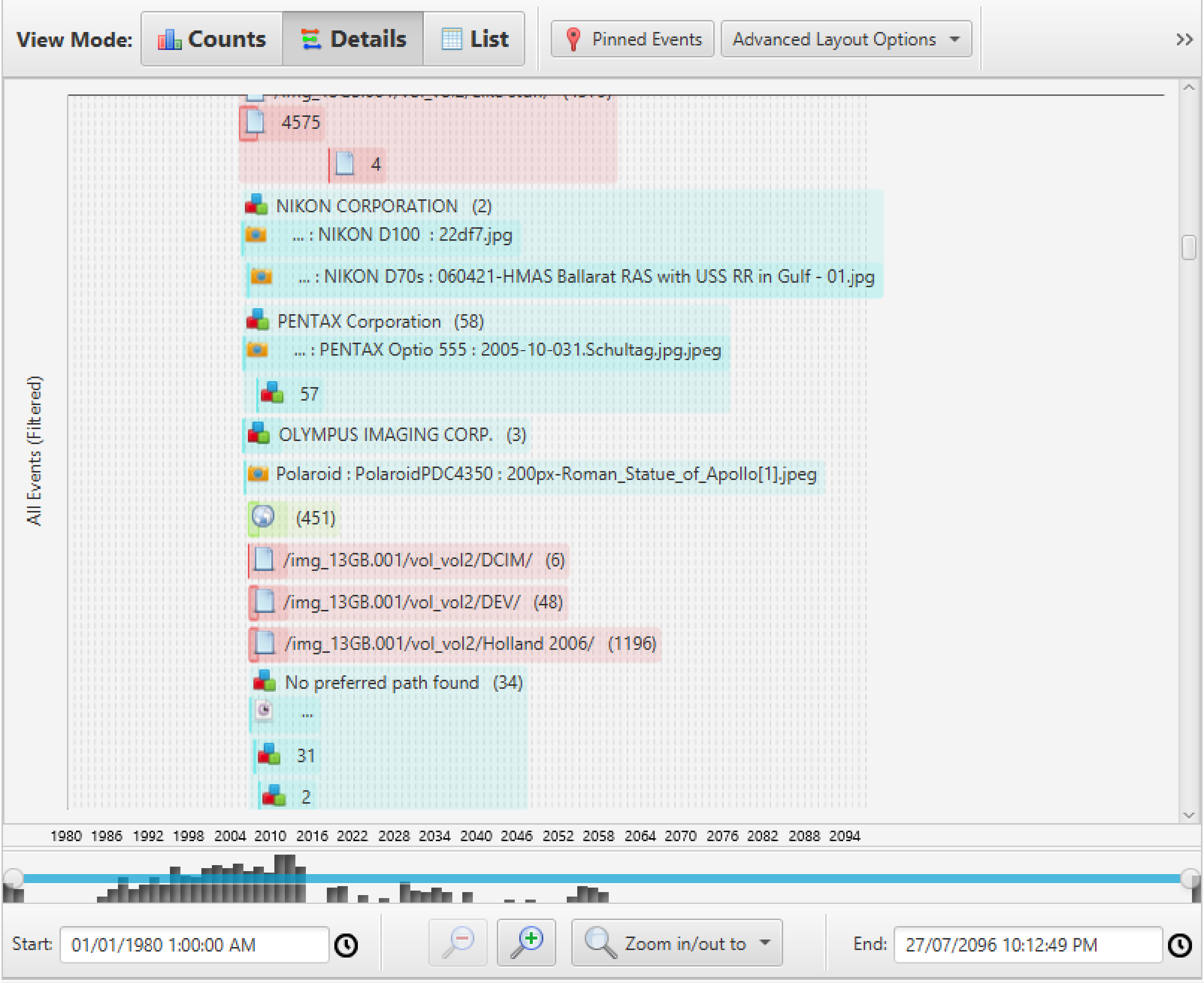
| | |
|---|---|
| Name | /img_13GB.001/vol_vol2/$OrphanFiles/--------.--- |
| Type | File System |
| MIME Type | application/octet-stream |
| Size | 11468800 |
| File Name Allocation | Unallocated |
| Metadata Allocation | Unallocated |
| Modified | 1981-01-10 01:06:02 ACDT |

Figure 32 - Timeline Counts View

As seen in Figure 32, by selecting an event in the timeline, the user is presented with both the “Results View” and the “Content View” below the timeline. This gives access to the file and all the generated data with it.

In this case there are anomalies within the timeline. This will likely occur in most cases. On the far right, in the year 2096, there are some miscellaneous events that have occurred. This is EXIF metadata generated for images taken with a digital camera. This camera had the incorrect date and time set in its properties, therefore, the timestamp on the image was generated incorrectly. Most anomalies that occur in timelines are the result of incorrect dates; however, they should not be ignored and treated any differently. It may be harder to work with such files as the provenance will be harder to determine, but the content of the file may still contain relevant and important information.

Determining provenance is possible given the amount of data available through Autopsy and how it is presented. Correlations can be made with images taken from the same camera, which may have correct creation dates. If there is enough of this information, it is possible to establish a time frame in which the image in question was taken.



4575

4

NIKON CORPORATION (2)

... : NIKON D100 : 22df7.jpg

... : NIKON D70s : 060421-HMAS Ballarat RAS with USS RR in Gulf - 01.jpg

PENTAX Corporation (58)

... : PENTAX Optio 555 : 2005-10-031.Schultag.jpg.jpeg

57

OLYMPUS IMAGING CORP. (3)

Polaroid : PolaroidPDC4350 : 200px-Roman_Statue_of_Apollo[1].jpeg

(451)

/img_13GB.001/vol_vol2/DCIM/ (6)

/img_13GB.001/vol_vol2/DEV/ (48)

/img_13GB.001/vol_vol2/Holland 2006/ (1196)

No preferred path found (34)

...

31

2

Start: 01/01/1980 1:00:00 AM

End: 27/07/2096 10:12:49 PM

Figure 33 - Timeline Details

The “Details” view in Figure 33 provides a lower-level perspective of the timeline. Rather having an overall picture of the events, this view offers a chronological ordered list that uses the same colour-coding as the “Counts View”. As the user scrolls down the view panel, the information displayed will move further along the date axis indicating the year of that event. Specific events can be selected, providing access to the results and content views.

The final view is a list which can be sorted by the date and time of an event, the event type, and the other fields such as the description which indicates the name or path of the file where the event occurred.

The three visualisations discussed are not the only way to visualise the data discovered through Autopsy. There is a reporting feature that allows different types of reports to be generated. The first two reports take the data from the results of the case and display them in either an interactive HTML report, or an MS Excel spreadsheet. A tab delimited text file containing information about each individual file can also be generated. A KML [Keyhole Markup Language] (OGC, 2019) format report with coordinates can be generated to use with Google earth views. A TSK (The Sleuth Kit) Body file can be generated that reports all the MAC (modification, access, change) times for every file and can be used to establish timelines. The last option is dedicated to forensics, known as STIX (Structured Threat Information eXpression).

## 6.2 Collection Institution Relevance

Showing the capabilities of available software and system configurations and their potential is the objective, rather than suggesting which specific tools should be used. Collection institutions ingest and preserve data which are obtained through various means, one of which is donations. Donated material, depending on the policies each institution has in place on what they accept, may come in various forms.

For example, obsolete media such as floppy disks, CDs, and computing systems that have been disposed of without regarding security risk and conducting the appropriate disposal procedures. The frequency of this occurrence, especially on a large scale, such as an enterprise disposing their obsolete assets, is reducing as awareness increases. Security awareness is increasing as there is more media coverage on the matter and a plethora of education sources. Whenever a serious security breach occurs that compromises users and organisations, there are several articles written and spread across social media.

Despite this, there will be instances where negligence and poor judgement occur where media is discarded without any thought, leading to several different means by which the media can travel.

Hypothetically, someone could come into possession of such media and having no use for it, pass it on or potentially sell it in a garage sale (home market). The new owner of this media may see potential interest by looking through the data from a novice perspective. They then may choose to donate this to a local collection institution. This is one of the examples discussed in Section 4.2.1 Ingest Scenarios.

Through the donation policies in place, the owner will communicate what they believe is on the media. The institution can then estimate the worth of the information to the collection and determine if it should be accepted. There are several variables that may determine the outcome or that require consideration. For example, the donated media contains information on an iconic public figure.

It would be unwise to simply take the overt data and add it to the collection without some form of investigation to determine the accuracy and provenance. This is increasingly important since the donor has accessed the material on their own device, meaning some changes may have occurred and their personal data could be exposed.

The collection institution will perform an investigation such as extracting metadata from the known files, but the unknown files also require investigation. One document, or a set of correspondence, can entirely change the nature of the original files. Most cases may not involve any controversy; however, the data may prove to be beneficial to the collection.

Furthermore, this study has revealed the potential amount of data that can be extracted from a small 13 GB hard drive that was used over 13 years ago. The use of technology has grown exponentially, made possible through the growth of storage technology and computing systems. Thirteen gigabytes may have seemed large several years ago, but now, modern computing systems rarely come with hard drives smaller than one terabyte, this now includes solid state drives which are starting to increase their standard size. With the tools and methods demonstrated on a small dataset, the potential discovery if used on modern devices, is exponentially larger.

The topics discussed in Chapter 4 AUSTRALIAN LAW IMPLICATIONS surrounding the types of metadata and sensitive data are significant given the capabilities of digital forensics combined with the nature of donated material. Metadata may still be a foreign concept to novice users. Therefore, when donation negotiations are conducted, if the interviewer understands the potential of the digital forensic tools at their disposal and the potential sensitive data residing on donated media, they can prevent or at least mitigate any legal and ethical issues that may arise. This is why the donor agreement stage of digital preservation is deemed essential in this study and is one of the focal points of the enhancements presented in Section 7.3 Workflow Enhancements and Visualisations.

As time progresses, there will be a shift in the general understanding of computing devices and how they are used, such as increased awareness of our digital footprint. It is becoming easier to be aware of such things, as the knowledge required is quite low. A novice user understands a VPN (Virtual Private Network) provides them with anonymity when searching the web, but they do not need to know how it works.

New generations of users are brought up with technology and are provided with more utilities that allow more applications to be used without requiring training. More usage creates a larger digital footprint. Therefore, in time, investigating donated material will require thorough examination as the risk of sensitive data will be greater.

There are still collection institutions with a backlog of legacy media, such as floppy disks. Eventually, current, and modern media will make up a large portion of collections. Technology will reach a point where hard drives are no longer created. Therefore, current media will become historical and in need of preservation.

If an iconic figure dies, the preservation of their information is important. Fame and infamy do not end with their passing and the knowledge of this figure, derived from the information collected, can determine how they are perceived by the public. It is up to collection institutions to take in this material and preserve it, ensuring the integrity and accuracy of the information whilst maintaining accessibility for future generations.

There is more to it than just preserving an artist’s work, for example. Once they have passed, others may wish to learn about their creative process, influences, routines, and other factors that may have contributed to the creation of their art. Digital forensics, specifically the two tools used in this chapter, Bulk_extractor and TSK, were used in a study involving the exploration of the late artist Stephen Dwoskin (1939-2012) (Bartliff et al., 2020). The findings of this study revealed that timeline and file (including metadata) analysis have the potential to reveal clues about the personal and professional history of that artist. This includes creative process, technical environment, technological choices, patterns of creative flow, and other elements that provide context surrounding the artist’s creations.

The discovery of new and sensitive data is not the only benefit digital forensic software and methods offer to collection institutions. As demonstrated, access to a gallery that can be sorted by the camera make and model, visualisations that show where and when events occurred on a system and being able to visualise communications are all beneficial when dealing with a complete system. Digital forensic software allows additional plugins to be installed, mainly for law enforcement purposes, but plugins such as video triage could prove useful. Video triage allows video footage to be broken down into frames to identify content easily and quickly within video files. This can be used to search for material such as sensitive content, overwrites, errors, and edits. Significant reductions in the time taken to establish the merits of video files without having to watch them is an obvious benefit.

The point of the experiments with Bulk_extractor and Autopsy was to show the potential of using such a tool, but not to suggest any particular tool be used over another. Whether the tools demonstrated are used or alternatives, it is strongly suggested that some form of advanced sensitive data retrieval is performed. Collection institutions should consider tools based on the need of their collection and the resources available. FTK (AccessData, 2018) and Encase (Guidance Software, 2018) are examples of alternatives to Autopsy. There are options in the form of complete packages, such as Autopsy, and there are environments made up of multiple tools, such as BitCurator. However, standalone tools may be better suited, potentially requiring less training and resources. Proprietary commercial products can often provide a specific feature desired by an institution; however, open-source solutions are readily available and are commonly used that offer the same functionality and the ability to modify based on the user’s needs.

## 6.3 Summary

With these experiments, the research questions regarding where improvements can be made in data gathering and how digital forensic tools and techniques can be implemented for this reason are addressed. Furthermore, as these tools and techniques open new risks due to sensitive and personal identifying data, the visual representation of the output of these tools can give insight and perspective on potential discoveries, which in turn will promote thought on how to address these data regarding legal and ethical matters. Should existing policies and procedures be in place for sensitive data, they may be revisited should the potential of sensitive data now exceed current preservation workflows due to the implementation of far greater data gathering capability.

To achieve this, the experiments within this chapter have demonstrated data gathering beyond manual methods. Text based software used for sensitive data discovery has shown a glimpse of hidden and revealing information that can be found on a hard drive. From credit card numbers, email addresses, websites visited, and all search phrases typed into a web browser. The potential of data hidden in obscurity has been demonstrated on a randomly selected donated data source. Whilst the purpose of such information is usually used for law enforcement, collection institutions can also make use of this to strengthen and enhance their collections, ensuring accuracy and validity.

The capabilities of Autopsy, the user interface for The Sleuth Kit, has been visually demonstrated to show the various features that collection institutions can make use of, such as visualising timelines, communications, and having access to a gallery that can sort images and videos by different groups. Report generation on the discovered data can be generated in several formats. The identification of system information and connected devices can be used to aid in environment emulation as well as how preserved media was is intended to be used, indicated by a unique peripheral device.

Discovering sensitive and unknown data within a collection institution’s custody should be a priority, not only for the collections benefit, but also in the event that information is contained on the media that could lead to an investigation that is in the best interest of the public. Sensitive data may already be a threat, lying dormant in storage. This will be a significant issue as time progresses as will the amount of born-digital data being ingested into collection institutions.

Digital forensic methods should ideally become part of the preservation process. Not all collection institutions are equal and may not require extensive sensitive data retrieval based on the types and volume of files they preserve. However, if the intake increases over time or changes, the necessity for such methods may change with it.

As there are varying maturity levels throughout the collection institutions of Australia and around the world. Those with higher levels of maturity are setting an example for those in their digital preservation infancy. Therefore, institutions that are utilising digital forensic methods should be transparent in their procedures so that others may follow. Institutions that are in the position to adopt such methods should be doing so and documenting the process so that others may follow their example.

The first step in achieving this is to enhance the existing digital preservation workflows, amending them to accommodate digital forensics, and to visualise with the processing and management of sensitive data. It is evident some existing digital preservation workflows do not address sensitive data, based on workflow diagrams, descriptions, and information gathered. Whilst this does not definitively indicate sensitive data retrieval is not conducted, as it may have been omitted from the workflow, there is no evidence, visually and form an external perspective, that it is handled.

The next chapter of this study focuses on the improvement of the digital preservation workflows for Australian institutions and how they can be enhanced with digital forensics and better sensitive data handling. Chapter 7 WORKFLOWS presents reviews based on existing workflows and discusses design and notation options. The enhancements created are visualised with consideration of limitation and choice.

# 7 WORKFLOWS

The following sections focus on workflow efficiency and reliability. Tools and methods used within workflows have been discussed, but the focus is now on the workflow design and the processes within. The visualisation of workflows is important as they allow users to better understand the flow of processes and how each system or user interacts with one another. By reviewing and following workflow diagrams, it may be easier to detect any issues and unhandled processes that may not have been obvious beforehand. With modern workflow designs, typically flowing from left to right and making use of swimlanes, discovering flaws is easier.

Workflow diagrams offer many benefits, both to the institution they belong to, and to any external parties reviewing the workflows for their own means. Collaborative learning is the main criterion considered when analysing existing workflows, seeking for transparency and accurate representation of each institution's workflow. Although the main purpose in collaborative learning is to aid external viewers, there are also benefits to the institutions themselves as they may identify flaws within their workflows once visualised.

An analysis and evaluation of each workflow that has been collected or derived from public and private information has been conducted. This allows the identification of where improvements can be made, good and bad design, and any standard notation.

There are two sets of workflows, the Australian and the U.S. Each set has been evaluated differently due to availability of visualisations, where the Australian workflow parameters have been extrapolated based on public information and questionnaire results. The workflows in each set have been evaluated against three criteria:

- Donor agreement
- Sensitive data discovery
- Sensitive data handling

As this study concerns the lifecycle of digital preservation, and not just the act of preserving, these criteria cover the processes from acquisition through to storage. The donor agreement check is there to ensure adequate information is gathered, where possible, from the donor as well as setting up appropriate measures in the event of sensitive data discovery. Sensitive data discovery refers to data passing through the workflow before being processed into a collection and whether it has been checked for sensitive data. How the workflow proceeds from there regards the handling of sensitive data.

The data collected from the participating Australian institutions did not contain completed workflows, nor diagrams, and were comprised of written examples and basic diagrams. In this evaluation, the data provided in each response helped in establishing a basic digital preservation workflow by correlating the information provided. This information included the steps in the preservation workflow, processes, and the tools and methods used. Although no dedicated workflow diagrams were provided, identifying potential flaws and where improvements could be made was possible. This was due to the supporting questions within the questionnaire and the additional data provided. Each tool used, and considered, as well as the processes surrounding these tools were described by the participants.

With all the data provided it was possible to determine if the specified criteria were met. The questions regarding donor agreements were straight-forward and covered instances where donors were no longer available, nor any next of kin. Much of this was supported where donor agreement documentation was made public per institution.

Sensitive data discovery was established by understanding how this criterion could be met with the tools and processes listed by each institution. If there was no mention of a tool or process that could perform these tasks, then it is probable these steps are not being performed based on the data provided. These data were correlated with the responses to the questionnaire questions that specifically asked about sensitive data discovery and the use of digital forensics.

The handling of sensitive data naturally follows its discovery; however, this can be performed in multiple stages throughout the preservation workflow. Without appropriate means to discover sensitive data, then it is assured that sensitive data will remain unhandled, leading to loss of information and increased risk, both ethical and legal.

The U.S dataset includes completed and publicly accessible workflows, allowing for a visual evaluation on their design, processes, and notation. Each workflow has been described and checked against the three listed criteria. Diagrams have been provided for each source of U.S workflows, revealing a total count of how many met the specified criteria. The exemplary workflows have been marked to indicate a quality they do not share with the other workflows.

When evaluating workflow diagrams, it is important to keep in mind that they may in fact not be an accurate representation of the digital preservation process. The personnel responsible for creating the workflow diagram may not have the experience needed to accurately visualise a complete process. They may be isolated from some of the digital preservation stages, meaning they do not possess all the required information to accurately visualise the workflow. Some steps may be omitted intentionally or by accident. However, regardless of whether these workflow diagrams are followed exactly; how they are presented; if they are publicly available; needs to be considered as there will be other institutions that use this material as a guide. Therefore, any public workflows addressed will be treated as if they are the complete representation of the workflow process. Data gathered from the Australian institutions derived from the correspondence from each participating institution in the questionnaire are more accurately attuned to the workflow process.

Sections 7.1 Workflow Evaluation - Australia and 7.2 U.S Collection Institutions - Workflows discuss the analysis of workflows within Australia and those collected from U.S collection institutions, extrapolating key information to guide the suggested enhancements.

Section 7.3 Workflow Enhancements and Visualisations presents workflow design and notation options, followed by workflow enhancements that have been designed to address the criteria used in the evaluations.

## 7.1 Workflow Evaluation - Australia

The following two sections focus on the evaluation of workflows, conducted in two parts, starting with the Australian set. Australian institutions were investigated by establishing what could be found in publicly available online material and creating a questionnaire based on what information was missing. The data gathering and analysis performed regarding U.S institutions aided in the development of the questionnaire. The U.S institutions, specifically U.S universities, offered more transparency in their publicly available material.

To begin with, the target of improvement was all institutions across the globe. This was soon realised to be an inefficient goal due to the ethical obligations and different jurisdictional boundaries. The institutions outside of Australian jurisdiction, their progress, and their workflows, provide useful information that can help achieve the goals within Australian collection institutions. This information can help establish a baseline and goals to work towards. The U.S institutions investigated are performing digital preservation at a higher maturity level which have adopted digital forensic methods to some extent, with varying influence across the institutions.

Comparisons are often made between two very similar datasets; however, in this case, contrasting what is, to what potentially could be, has provided desirable information. The two datasets provided different avenues to explore regarding tools used, the overall preservation process, and workflow design and notation where diagrams were present.

### 7.1.1 Donor Agreements, Sensitive Data

Prior to the discussion of workflows, it is important to address the procedures and policies in place regarding donated material, typically addressed within donor agreements. Since the discovery and handling of confidential and sensitive data is the primary area of focus, achieved via adopting digital forensic tools and methods, the agreements made between donor and recipient are the first stages for the handling of such data. Should this not be conducted adequately, the entire process could be compromised.

Although it is important to make the decisions that establish what happens to the data once ingested and processed, it is also about extracting as much information regarding the donated material from the donor. The information elicited from the donor may be crucial in giving context to any discovered sensitive material. This information will expose any ties the donor may have to the material, which could, in the event of illicit content discovery, exonerate or condemn the donor. This information may not always be available as the source of the material is not always known or divulged and the donor may only be a custodian of the material, with no ties to it.

The questionnaire provided to the institutions of Australia, (Appendix B - Questionnaire), begins with questions regarding donor agreements and ethical standards. These questions aimed to determine if proper donor agreements were in place and to determine how cases where handled that were not covered in these agreements. This included when the donor was no longer available, and the decision had to be determined internally.

The responses from the state archives that participated in the questionnaire will not be included nor counted towards any of the results provided within this chapter as the information provided was incomplete.

Question one addressed the processes in which ownership, access, and donor stipulations are handled. Each institution had their own stipulations and criteria, but each one allowed donors to stipulate how their donated material is handled. Amongst the institutions that were approached, donor stipulations can often be negotiated; however, derived from public information, one institution does not accept donations if the donor has any conditions and does not relinquish full access and ownership of their material. As with most policies, exemptions do apply. For consistency, anonymity of institution names, locations, identifying factors and actual responses will be maintained for institutions within Australia.

The variations based on the question's responses include:

- The donor may stipulate ownership and access conditions.
- All conditions must be communicated and agreed upon before material is accepted.
- Negotiations take place, but all restrictions must have an end date and the positives must outweigh the negatives.
- A highly detailed deed of gift must be completed, allowing all stipulations.
- Provenance and relationship with material is discussed. Copyright ownership is recorded for all holders (family members, etc.). Legal conditions and stipulations are agreed upon. Sensitive and cultural content embargoed must include an end date by the donor.

The final point in the list above addresses the concerns expressed and is exactly how the donor agreement process should be conducted. All areas, no matter how trivial they may seem for some cases, must be addressed, preventing any issues that may arise. Preventive measures are always better than mitigating the damage once an event occurs.

The questions that followed addressed the discovery of sensitive data that:

- has been addressed in the donor agreement
- has not been addressed in the donor agreement
- has no next of kin or contactable owner.

Whilst the first question may seem self-explanatory as one can assume the answers would consist of "follow the agreement", there are additional details requiring discussion. Each response that differed from the others is briefly described. Regarding the first question, the discovery of sensitive data that **has** been addressed by a donor agreement:

Among the institutions, issues regarding sensitive data have been unprecedented. In one instance, sensitive content is made accessible, uncensored, with the option to place warnings and locks on the content, making it accessible onsite only on selected terminals within the institution. In this response, if the content is covered by the agreement, it is not necessary to further involve the donor.

When dealing with donor specified stipulations, if reasonable, are enacted per the donor's request. All restrictions are reported and are then reflected in catalogue records. Access rights are then determined based on the requirements as material is ingested into the collection.

Other ways in which stipulations are recorded include embargos and access conditions to be embedded in descriptive and administrative metadata, assuming this means the institution will adhere to these conditions. However, if the material discovered, even if covered by the agreement, is significant enough, the donor agreement may be revisited, involving re-negotiation with the donor if available.

There are simpler instances, especially within smaller institutions, where adhering to stipulations according to the agreement has sufficed. Several levels of access are offered to meet these requirements and embargoed content is honoured per request of the donor.

The infancy of the situation has been addressed by one of the participants. In providing access to digital material, it was stated there have not been any issues regarding sensitive material and donor agreements. The institution is taking preliminary action in discussing the topic as they are aware that legal agreements need to be updated to accommodate this topic.

These types of institutions are willing to accept materials accompanied by donor specified stipulations; however, there are differing levels of flexibility among them. Some institutions strictly follow the agreements, whereas others are willing to re-negotiate when needed. The level of stipulations accepted also differs among the institutions.

Many of the issues discussed throughout this study regarding sensitive data have not surfaced for all the institutions. One must consider if the unprecedented issues have not surfaced due to the means by which to discover sensitive data are not part of the digital preservation workflow.

The next question regards situations **not** covered in donor agreements. The results included short responses and some in-depth answers. A situation occurring that is not covered by the donor agreements is unprecedented for each of the institutions that responded. At the time, preparations were being made for these situations, acknowledging the potential risk and inevitability. The likelihood of it happening increases with time as the born-digital age progresses, and with it, the volume of intake as more of our digital history requires preservation.

Two of the responses clearly state this is an unprecedented occurrence and the donor will be addressed. There was, however, a more informative response from an institution that has prepared for such an event, providing a detailed action response. In the event of sensitive material discovery, the acquisition process is halted, the donor is contacted, and a new agreement is made. This may result in sensitive material being returned or documented appropriately. Instructions are in place to help mitigate these situations by instructing the donor to empty their trash, email, accounts, and any other personal information if the donated material belongs to them or has been accessed on their personal systems.

This shows a clear understanding of the potential threat of sensitive data and the sources from which it comes. In this specific action plan, asking the user to clear their personal data, indicates the institution is aware that donated material may originate from or has been accessed on the donor’s personal computing device. Accessing data may create or modify metadata. This is a necessary precaution, and one that should be shared for all institutions.

Another response also provided some additional detail. Although an unprecedented occurrence, there are policies in place covering a range of different cases. Firstly, legal obligations are considered such as privacy, defamation, and copyright, addressing the exemption from such laws for the collections in Australia’s jurisdiction. They are still followed as guidelines.

Specific protocols are followed as needed, such as the ATSILIRN protocols should the content be related to Aboriginal and Torres Strait Islander culture. The usual response here is to restrict access. No edits or deletions occur in heritage collections unless absolutely necessary. Other non-legal cases, such as public relations and relationship management concerns are considered grey areas and are taken case by case. Censorship and editorial control over collection material is not exercised. As this has not been an issue yet, specific protocols for such an event have not been developed.

One institution also claims this is yet to be an issue and that it would be handled by addressing the donor and re-negotiating a new agreement. They do, however, have protocols in place for non-digital material and would enforce these protocols as there are none specifically addressing this situation.

The responses given provide a detailed insight on how some of the institutions address and foresee the potential issues surrounding such events. Some still regard the laws that surround privacy, even though their collections are exempt, a practice that needs to be enforced throughout all institutions conducting digital preservation. Specific policies are not in place yet, but existing ones are considered to help make decisions and handle these situations.

The question regarding donor agreements and sensitive data discovery is based on the ethical procedures in place should the owner nor next of kin not be available to contact.

One of the institutions indicated, although an unprecedented event, take down procedures are in place in accordance with the guidelines from the National and State Libraries Australia (NSLA). For the removal of content to take place, the following considerations are to be made on whether:

- the material is defamatory under Australian law, or is subject to a suppression order
- the material contains sensitive information about someone who is still alive, and making that information publicly available online puts them at risk of serious harm
- access is consistent with copyright law
- access is consistent with wishes expressed by the donor
- access is consistent with the guidelines for collaborative practice between Libraries and Aboriginal and Torres Strait Islander Communities
- taking down the material would contradict democratic principles of unrestricted access to information and ideas.

Among the other participating institutions, procedures include consulting a sensitive collections policy and taking care not to make material accessible until the following considerations are made: Creator's moral rights, implications, standard time frames for making restricted material available, and consultation with legal advisers.

Content is usually made accessible if the original agreement allows it in good faith; however, legal considerations are taken first if required, in the event of a potential defamation case, for example.

Whilst an unprecedented event, there are institutions that still identify the risks if such an event where to occur and act accordingly. One institution employs preventive measures which are taken whilst still in contact with the donor, attempting to discover any issues early to reduce the chance of such a case happening where the donor is no longer available. If this were to occur, the content would likely not be made public. No redaction is said to have taken place to date within the institution's collection.

One of the respondents provided some additional information:

> *"We would make a decision case by case, depending on the material in hand and the Library's Copyright determination and risk management framework."*

> *“As noted above, the archivists will impose restrictions on any material which we consider inappropriate for general access (mostly records relating to people who can be presumed to be still alive) even if the donor has not imposed any restrictions. In the case of physical items we withdraw them from the general collection and note in the bibliographic record that access is restricted, who may access the records and under what conditions.”*

How these issues are managed is still under consideration and seen as a priority. An example was provided with the response and is summarised as follows:

A collection of records was accepted based on crime victims. It was assumed only the business records were collected; however, further processing revealed confidential notes with identifying information. The records could not be returned as the entity no longer existed. Advice was sought re the Privacy Act and it was determined to embargo the records for 70 years.

From the information gathered on all the questionnaire responses regarding donor agreements and sensitive information, it is clear that whilst somewhat inconsistent and not entirely concrete, some form of sensitive data handling is being conducted. However, this information combined with the remaining questionnaire results, specifically on the tools and processes used, indicates that sensitive data handling could not be performed efficiently nor accurately due to the lack of tools and procedures that make this possible. This applies to all institutions, although some have taken the first steps by owning equipment and tools that are used for this purpose, though at the time, were not in use. Without the appropriate tools and procedures to scan and identify, much of the sensitive data that may be present on donated material will remain hidden, therefore, cannot be processed in accordance with the guidelines.

A recent study provided a description of the work being conducted by a volunteer at a state archive in the United States of America (LeClere, 2019). This was centred around the digitisation of materials, but although the material was not born-digital, the experience is relevant.

Every document needed to be scanned, indexed, and tagged. This involved indexing every name found within the documents. The archive’s volunteer was determined to get every name. The reason being is that discoveries were made on the lesser-known people that were in fact far more important than many of the bigger, well-known names. The need for these lesser-known people to be recognised for their accomplishments was important. Some of the collection items being digitised were interviews for voluntary positions with Civil Rights organisations. The focus was on the volunteers that were part of the project. In the thorough investigation, it was discovered that the individuals rejected were also within these documents. The volunteer suggested this information not be digitised and added to the collection for it was not relevant and it contained "pretty damming or damaging comments" about the rejected individuals, their mental capabilities, and their opportunistic reasons for volunteering.

Further key findings from this study include the challenges of resources required such as time, money, and labour to support large-scale digitisation projects and how much of the funding is needed from outside sources and grants, said to be the "lifeblood of most digital projects".

The experience of the volunteer brings to light the challenges faced in large-scale digitisation projects. Whilst this cannot directly be compared to the experiences of some of the smaller institutions here in Australia, one can assume there are similar challenges. Smaller projects, less resources, and whilst the preservation process is a little different for born-digital material, challenges still exist. The identification of the sensitive material and the recognition of lesser-known individuals deserving of credit allows the institution to give credit where it is due and to save certain individuals having defamatory information revealed about them for no justifiable reason. Whilst this was an onerous task for the volunteer, some of the effort can be alleviated for born-digital content as there are tools that aid these tasks. Whilst tools exist to discover, sort, and present sensitive data in a meaningful way, the human analysis still plays a pivotal role.

Although considerations are made regarding sensitive data, improvements and standardisation across the institutions is required and should be desired across each institution. Whilst some institutions may not acknowledge the necessity in adopting the means to handle sensitive data, due to their intake volumes and experience, the risk should not be overlooked.

Another concern is how exactly is this sensitive data being discovered? There are no indications within the workflow examples provided that specifically handle this process, therefore it cannot be confirmed it is being done. There may be instances where one may run a search within a directory and deem this as adequate sensitive data retrieval. The experiments and findings in this study have already proven this to be inadequate. Without the aid of digital forensic tools, significant amounts of information will remain hidden.

The lack of details regarding sensitive data handling within workflows published by institutions outside of Australia is concerning. Alternatively, some workflows indicate the use of dedicated tools to support sensitive data handling; however, the workflows do not visually indicate the processes involved. For example, when following data through a workflow, once it reaches a node dedicated to sensitive data discovery, the data proceeds to the next node regardless of the outcome. Therefore, no decision-making is being visualised. Ideally, there should be various outcomes that are determined by the results of the sensitive data discovery process. Furthermore, there are no subsequent nodes in the workflow that indicate what happens to the data.

It is understood that workflows do not necessarily visualise the complete process; however, as it has already been argued, transparency is crucial in this field as institutions all over the world are at different maturity levels of digital preservation and are looking to their peers for guidance. Furthermore, regarding the U.S institution workflows, it makes no sense for information regarding digital forensic based methods and processes to be omitted. This is especially true for institutions that are clearly influenced by digital forensics.

### 7.1.2 Workflow Extrapolation

Establishing a definitive workflow for each institution participating in the questionnaire was not possible for several reasons. Some institutions were not mature enough in their development and were not using a dedicated digital preservation workflow, whereas others are in the process of implementing new procedures and tools. Smaller institutions, based on their intake, are adopting procedures as needed, case by case.

Three workflow examples were provided, two of which were a list of steps in the digital preservation process, one of which was an assumed set of steps based on the OAIS model as the institution at the time was in an implementation phase. The other example provided was a diagram resembling a basic flowchart, not developed enough to be considered a workflow diagram. One institution only has procedures in place for ingesting and file naming with no dedicated digital preservation plan.

The following list, derived from the questionnaire results, provides an overview of the core processes conducted collectively among the participating institutions:

- Write blocking – not all the institutions that responded utilise this crucial step in their workflow.

- Checksums (fixity) – this was only used in special cases and not often for one of the institutions, the rest used it in all cases.
- Descriptive metadata – additional metadata were not added within the preservation process for one of the institutions, the others put great emphasis on this process.
- Disk imaging – one institution clearly states the use of disk images in the response to questions about workflows and tools to facilitate their preservation process. The other institutions do not claim to utilise disk images; however, one of them has said to use them only in special cases.
    - It is understood not all files require disk imaging, this could be a result of how material is delivered to the institution and does not necessarily reflect on the institution's process.
    - When asked about any digital forensic software and hardware, all responses suggest FTK imager, which was not reflected in earlier answers. There are some responses that indicate the digital forensic tools listed are only used in special cases or have just been acquired at the time.
- Access permissions – Applied in some from across all responding institutions
- Preservation packages, Library Management System (LMS), asset management, etc.
    - Dedicated preservation tools were not in use across all institutions, but a form of asset management was being conducted.

From the data gathered, the extent of forensic utility includes the creation of disk images, write blockers, and checksums. This was evident in the answers provided regarding workflows and digital forensic tools in use. Furthermore, some institutions stated the need for improvements to their workflow and one institution was clear they did not want any changes. The reason given was the introduction of new tools would require the training of archivist staff. The resources needed based on the volumes of material needing preservation are factors that influence responses such as this. It is hoped that should the need for preservation grow, more resources will be made available for the departments responsible for such actions. Therefore, institutions developing their preservation practices later than others are likely to learn from further developed institutions. This makes workflows, workflow improvements, and all associated studies relevant to them should futureproofing be considered worthwhile.

Regarding resources and the preservation needs of each institution, the tools in place may be adequate for their current workload which is why changes to the core preservation process are not suggested. Improving and visualising sensitive data discovery and handling is an amendment that does not need to affect the existing tools nor have any significant impact on procedures. The additional processes will require the investment of more time, and potentially more staff, a trade off that if accepted, will elevate the overall preservation workflow. Training costs may be involved; however, the investment to reduce future risk and enhance the data gathering capabilities of the institution should be carefully considered.

Whilst no changes are being suggested to the core of existing digital preservation workflows, this does not indicate improvements cannot be made. Standardisation and consistency could certainly be improved as has been shown in the inconsistency in tools used across institutions. However, the goal presented in this thesis is to prevent future issues from the hidden threats within born-digital collections. To achieve such a goal, it is necessary that workflows reflect the events that occur after the discovery of sensitive data and the tools and methods used to handle the discovery. If only one part of a collection is to be made public whilst the rest remains in storage, as it is not being used, there is no harm in processing it with digital forensic software. It will, however, take time and resources. It is understandable if the resources are not available, especially in small institutions, but it is strongly suggested that it should be a consideration because the media on which the material is received, in the case of a donation, could contain information with various consequent impact factors.

The next section discusses workflows presented in the datasets collected from the U.S. The results in this section require less interpretation as there are definitive workflows to assess. However, some extrapolation is needed as some workflows can be vague.

## 7.2 U.S Collection Institutions - Workflows

The data gathered in this section were easier to collect as the information provided to the public is much more in depth than what could be found without communication with the Australian institutions. The collection of workflows gathered resides in two sets. The first set has already been analysed for the tools used, that is the 2012-2016 dataset from the BitCurator Consortium. These will be briefly re-analysed, focusing on sensitive data detection and handling, as well as any visualisation of donor agreements being conducted.

The second set comes from the Educopia Institute Community Cultivators project, OSSArcFlow - Investigating, Synchronising, and Modelling a Range of Archival Workflows for Born-Digital Content (Educopia, 2018; Post et al., 2019). This set contains 12 current and in-use workflows. Each of these workflows uses the same three tools: ArchiveSpace, Archivematica, and BitCurator. The focus is on the processes and the flow of events; therefore, this dataset has not been included in Chapter 5 WORKFLOW TOOLS – DATA GATHERING as it is better suited to the topic of this section.

The following criteria are believed to be what is commonly lacking in workflows and in need of change.

Event handling - events that are not handled properly, or do not accurately represent what would occur in these situations. An example of this is if sensitive data discovery is visualised in a node within the workflow, are there steps involved to indicate what happens if there is a discovery, then what happens to that discovered material? In the event that sensitive data discovery occurs, what steps are in place to halt the workflow or allow it to continue once the event is handled?

There may also be a lack of handling, meaning there is no visualisation of decision-making when events occur, representing a workflow that continues regardless of what is discovered along the processing path.

A major concern is if sensitive data discovery is not being conducted. This poses a serious threat, even if precautions are in place that restrict access to the material once preserved in a collection, threats can emerge from inside the institution.

The last criterion is focused on donor agreements and how they are visualised as well as any events that surround this process such as accession record keeping. It is beneficial to visualise such events as it may indicate flaws in the process.

### 7.2.1 Workflow Analysis

For the analysis of workflows, anonymisation was applied to remain consistent. Each institution has been given a unique identification number. These IDs range from MEM1 to MEM12 for the BitCurator Consortium workflows and OSS1 – OSS12 for the OSSArcFlow workflows. MEM1 through to MEM7 are the workflows from the 2012 dataset and the remaining are from the 2016 dataset. The OSSArcFlow dataset are listed as current (2018-2019). Different IDs were used to differentiate the two sources of workflows to allow for comparisons and the identification of shared or unique variables.

A short description for each institution is provided accompanied with visualisations of the two sets, separate and combined, to show how each of the criteria is met. That is, donor agreements, sensitive data discovery, and sensitive data handling. The exemplary models are distinguished and discussed further as they provide potential instruction in determining the best approach to apply the suggested enhancements.

### BitCurator Consortium Workflows

The unique identifiers **MEM1** to **MEM7** are the institutions from the **2012** dataset. **MEM8** to **MEM12** are from the **2016** dataset.

#### MEM1

The initial consultation with the donor is reflected in the workflow, along with the analysis and record keeping of the acquisition and accession of the collection. High priority material may be processed in FTK, but there is no indication the purpose is sensitive data retrieval. There are no decision-based nodes that reflect how sensitive data would be handled, instead an access copy is created for distribution to reading rooms.

#### MEM2

The workflow represents the initial donation and acquisition, followed by accession record keeping. There is a clear indication that checks for sensitive data are performed and reported before the processing of access takes place. This represents a form of handling. There is a pathway through the workflow where if a reader is not available, the media is stored in a database with no further action taken. This poses a security risk if, as indicated, the media does not pass through the sensitive data check. If the institution were capable of reading said material, it should proceed through the appropriate checks.

#### MEM3

The initial donor agreement is present along with basic record keeping of inventory. There is no sensitive data discovery or handling. There is a decision node which reflects the type of content being processed, one of which is a computing system, which is then used on-site as an access point. As no sensitive data discovery or handling is indicated, this poses a security risk of uncertain severity as the access restrictions are not clear.

#### MEM4

The donor agreement and following nodes within the workflow indicate an initial check and appraisal, including a log of media data and metadata. A node exists within the workflow that contains the process for a secondary appraisal and re-image if necessary. Removal or redaction of information takes place in this step, but it is not clear how this process is handled, and it is not visualised. This raises concern as the removal and redaction may be based on data discovered via manual search methods without the aid of digital forensic tools, meaning only a small fraction of data has been revealed.

**MEM5**

Initial processes include donor acquisition and record keeping. Sensitive data discovery is indicated within the workflow and includes redaction if necessary. This node is made up of multiple processes which visualise the flow of data from the creation of the disk image, quarantine protocols, and the discovery and redaction of sensitive data. This process is handled with dedicated hardware and software, implying the sensitive data is handled correctly. The final terminating node in the workflow indicates access mechanisms are investigated, which means throughout this workflow, appropriate efforts are made for sensitive data discovery and handling. Questions remain about the redaction process and the handling of the original data.

**MEM6**

See Mem1. Whilst small differences are present, the workflow is very similar as it comes from a different department in the same institution, handling different media.

**MEM7**

The acquisition method is different from previous workflows and involves an initial analysis of the recordkeeping systems in place by the donating party. There is no visual presentation of any agreements made. No sensitive data discovery or handling is represented in this workflow.

**MEM8**

MEM8 is the first workflow from the 2016 dataset (MEM8 – MEM12) which shows considerable change compared to the 2012 dataset (MEM1 – MEM7). The process of the donor agreement is expanded on further, showing the decision process if media cannot be processed by exploring alternative means. The discussion of privacy and access issues with the donor are included as a step in the workflow. Whilst not a focus point, the virus check within this workflow is handled by a decision to terminate and not proceed in the event of virus discovery. Decision-making such as this is sought after as the process is halted when necessary, instead of continuing to process regardless of the discovery, an issue in some workflows. Sensitive data discovery is implied as “PII reports” are generated. This is followed by groupings and access in the adjacent node where there is an extraction process which involves the mounting of the disk image to extract and normalise objects. Beyond the reports on sensitive data, there is no visual indication as to what happens to these data.

**MEM9**

Upon acquisition, accession records are created or updated. There is no visualisation of donor agreements. A decision node is in place to determine if the contents of a disk image require a search for sensitive data. If yes, a search using an appropriate tool is conducted; however, the next node in the workflow is proceeded regardless of the outcome. There is no indication of what takes place when sensitive data are discovered, therefore, it is not being considered as handled.

**MEM10**

The donor communications and discussions are represented in this workflow. Sensitive data discovery is performed, but there is no indication on how it is handled. If the collection is to be made public, an access copy is created and there is an instance of the sensitive data discovery software being used in the final node of the workflow, but it is unclear how.

**MEM11**

The donor agreement contains more steps than previous workflows have included. These nodes represent the negotiation that takes place and the decisions that need to be made before acquiring the media. If the content is digital, an interview with the donor is conducted. Once the acquisition is accepted, a donor agreement is created, followed by an accession record. Sensitive data discovery is performed with the creation of a disk image. Before access is provided, the sensitive data reports are reviewed along with the files. There is no visual representation on how this is handled, nor any decisions made based on these reviews.

**MEM12**

Donor negotiations and agreements are presented as nodes within this workflow. Scans for sensitive data are performed, but there is no visualisation or decision-making once done so. The workflow proceeds each step, regardless of outcomes.

Figure 34 illustrates how each of the criteria was met across the 12 BitCurator Consortium institutions. This shows that only 5 of the institutions were handling sensitive data effectively; 8 performed sensitive data discovery; 11 had a donor agreement process.

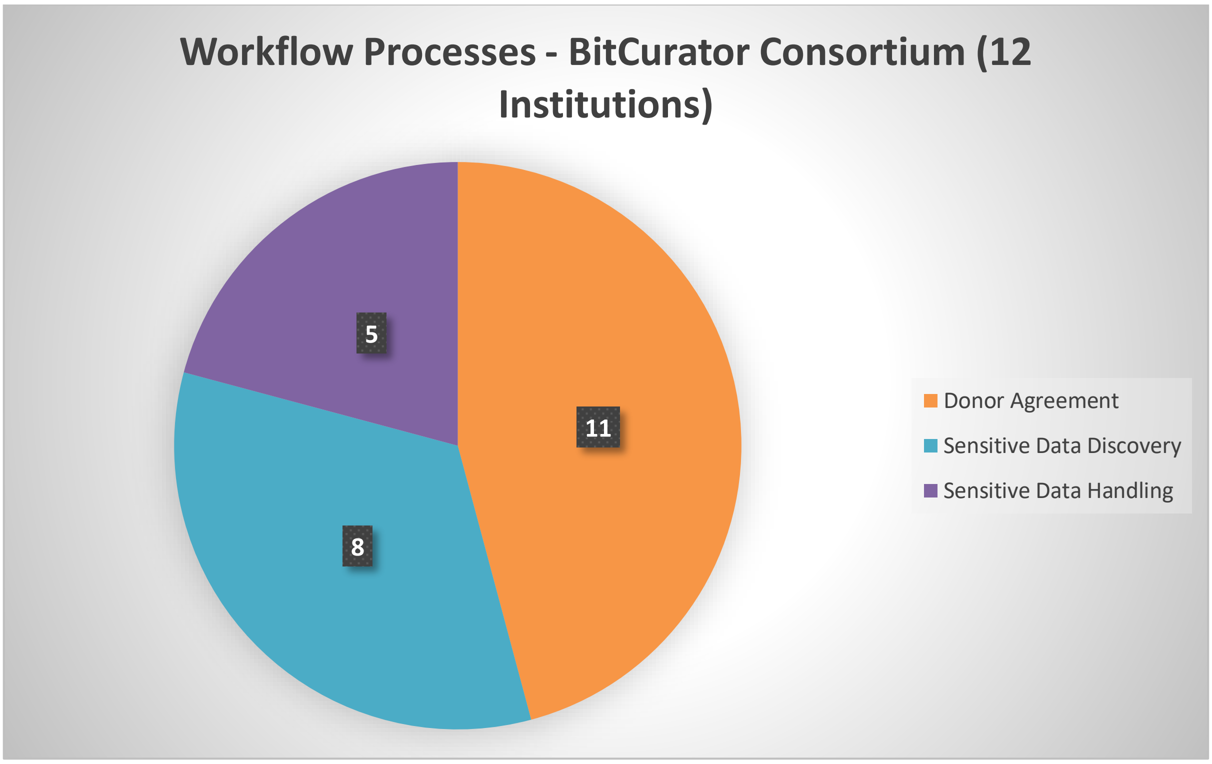


Figure 34 - BitCurator Consortium – Workflow Criteria Breakdown

### OSSArcFlow Workflows

**OSS1** to **OSS12** represent the **2018-2019** OSSArcFlow dataset. The workflow diagram format allows the institution to indicate the use of ArchiveSpace, Archivematica, and BitCurator. An example is provided in section 7.3.1 Notation and Design, Figure 42 - Alternative Swimlane (Software/Tools). Within the following reviews, "*" indicates an exemplary workflow and "**" indicates an exceptional workflow.

### OSS1

This workflow focuses primarily on the core process and does not display any instances of donor agreements, sensitive data discovery or handling. It does, however, make use of decision-making and appropriate UML notation.

### OSS2

The donor agreement and accession processes are done before any further processing takes place, with extensive record keeping and documentation, conducted by different levels of staff (Digital Curation Librarian, Processing Archivist, Library Staff). Metadata are retrieved and documented at various stages as are reports; however, there is no direct and clear node stating the scanning of sensitive material. Access copies are created along with rights and permissions metadata, but again, there is no clear handling of sensitive data if and when it is discovered.

**OSS3 ***

This workflow does visualise the donor interview process and accession record keeping before processing takes place. Sensitive data discovery and handling is done. Within the workflow, there are multiple paths, for example, physical media and logical files follow different routes throughout the workflow. They both reach separate clusters of nodes that scan for viruses and sensitive data before proceeding. The node that indicates the uploading of the submission information package (SIP) to the repository states that SIPs with confirmed sensitive data will be sent to a different storage location to comply with policy.

Overall, this workflow can serve as the exemplar due to its extensiveness and coverage of processes. One recommendation on an amendment that could be made is to visualise the storage of the sensitive media and the actions that follow. As the process to discover the sensitive data is not described, there may be more thorough methods that could be conducted leading to the discovery of more data with potential to impact the collection.

**OSS4 ***

The design of this workflow differs from the others as it uses a tier-based system. It is quite extensive, starting with the appropriate donor interviews and accession record keeping. There are multiple levels of decision-making which helps assign a tier to the content. Known formats with low sensitive data risk are put in a different tier to that of unknown or unusual high-risk formats. If sensitive data is expected, an analysis is performed, followed by a redaction or separation of files. Reports on sensitive data are reviewed before the content is packaged. It is still unclear what happens to the discovered data and if there is any further processing involved as the level of scrutiny is uncertain. However, the extent exceeds that of OSS3, making it another candidate to serve as an exemplar.

**OSS5**

This workflow follows three different types of ingest material. Donor agreements, interviews and accession decision-making are performed where necessary. The different pathways for each type of material ingested are visualised. There is, however, no visual indication of sensitive data discovery or handling.

**OSS6**

Material is carefully analysed throughout the workflow, starting with reviewing material with the donor or creator. Accession records are created, and the material is examined. Sensitive data discovery is performed with additional investigation into emails and faculty papers for sensitive and personal information. There is no indication how any sensitive data is handled.

**OSS7**

The workflow is of adequate design with appropriate decision-making, but there is no indication of events surrounding donor agreements, there is, however, accession record keeping. There is no mention of sensitive data discovery or handling within the workflow.

**OSS8**

Basic donor and accession record keeping takes place in the beginning of this workflow. There is no clear indication of sensitive data discovery or handling, but there is mention of automated report generation with FTK. How FTK is used within this workflow is unclear and sensitive data discovery may in fact be in use, but it is not visualised.

**OSS9**

Extensive analysis is performed before accession takes place, including the involvement of the donor. A decision is made based on where the material comes from. If the material is heterogenous or donated, it is analysed for sensitive data and a forensic report is generated. The material is reviewed multiple times before moving forward to reading rooms. The events that occur should sensitive data be discovered are not visualised. The fact that material coming from known sources is not processed in the same way as the rest is somewhat concerning. Although this may be a trusted source, there is still risk as the source may not have conducted the appropriate analysis before handing over the material. Unless there is full transparency regarding the provenance and change history of the material, there are risks involved by not processing this material in the appropriate manner.

**OSS10**

This workflow contains the initial donor involvement and accession record keeping. Regarding sensitive data discovery, this is performed only on physical media. The content is analysed for sensitive material, followed by a decision node that visualises content to be restricted in the event sensitive data is found. Whilst basic, some form of handling is being visualised. As to why material transferred over a network or destined for a public repository is not processed the same way is unclear.  If there are means to ensure this content does not contain sensitive material, it is not visualised within this workflow.

**OSS11 ****

This workflow is the prime exemplar. The donor process is extensive, including meeting or interviewing donors, gathering information, obtaining samples, and setting expectations of capabilities, all of which are visualised.  Material that is part of a processing project is immediately analysed for sensitive content. If no sensitive content is found, it proceeds through the normal workflow process. If discovered, it is transferred to secure storage before further processing takes place. Material that is not part of a processing project is also sent directly to this secure storage for later processing. If this material is requested to be accessed, it is reviewed for sensitive material once more and the screened material is copied for access in reading rooms. This provides an extra layer of security and serves as a quality handling of sensitive data.

Visualisation on what happens to the data in the secure storage would provide a good level of added detail, such as access restrictions imposed and if any further contact with the donor is conducted regarding the sensitive data.

**OSS12**

This workflow is quite extensive and highlights areas where there are issues, known as pain points. Extensive discussion is conducted with donors, determining all the characteristics of the donated material. In this stage, sensitive material is discussed in an attempt to identify it early, or at least determine the likelihood of its existence. Interestingly, the design of the workflow is negotiated within the workflow, meaning that whilst this workflow serves as an overall process scheme, the structure and processes are flexible. Access restrictions are applied upon restricted access requests. Analysis for sensitive data occurs, with a pain point stating the number of false positives makes it difficult as the reports must be manually reviewed, taking significant time. Within this set of processes, after the reports are generated and reviewed, remediation takes place, but there are no indications as to what this includes. How this is performed is unclear, but a remedy must take place for the workflow to continue as there is no visualisation indicating otherwise. The discovery of sensitive data is performed as is the handling of such data.

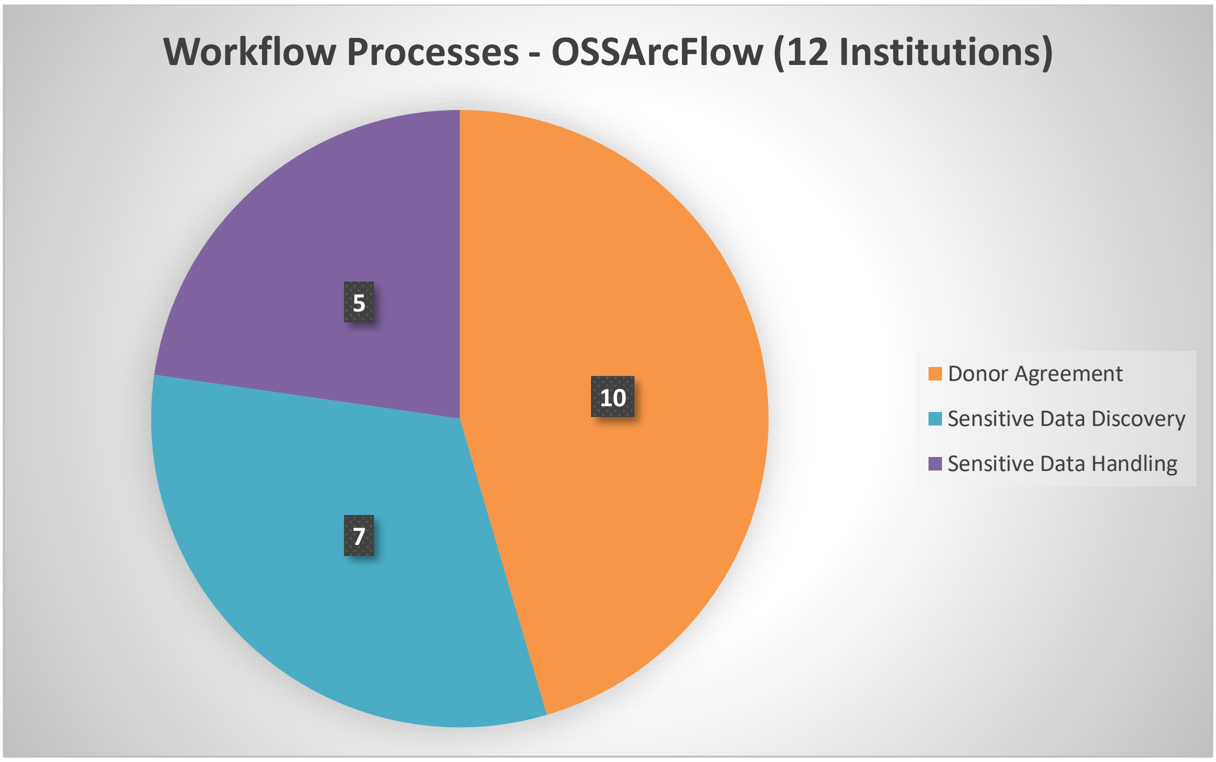


Figure 35 – OSSArcFlow – Workflow Criteria Breakdown

Figure 35 illustrates how each of the criteria was met across the 12 OSSArcFlow institutions. This shows that only 5 of the institutions were handling sensitive data effectively; 7 performed sensitive data discovery; 10 had an adequate donor agreement process.

In total, 10 out of the 24 institutions met all 3 criteria. Three of the workflows had uncertainties regarding the criteria being met. There was some indication that the criteria had been met or attempted, but the extent was unclear. Out of these three workflows, the uncertainties were as follows:

- Donor agreement – 1
- Sensitive data discovery – 1
- Sensitive data handling – 3

This indicates out of the three workflows with uncertainties, the sensitive data handling was shared across each of them. One of the workflows had uncertainties in both sensitive data discovery and handling. The workflow that had an uncertainty regarding the donor agreement also failed to meet the remaining criteria. Two of the 24 workflows failed to meet any criteria.

The following chart shows the combination of both datasets and how many criteria were met:

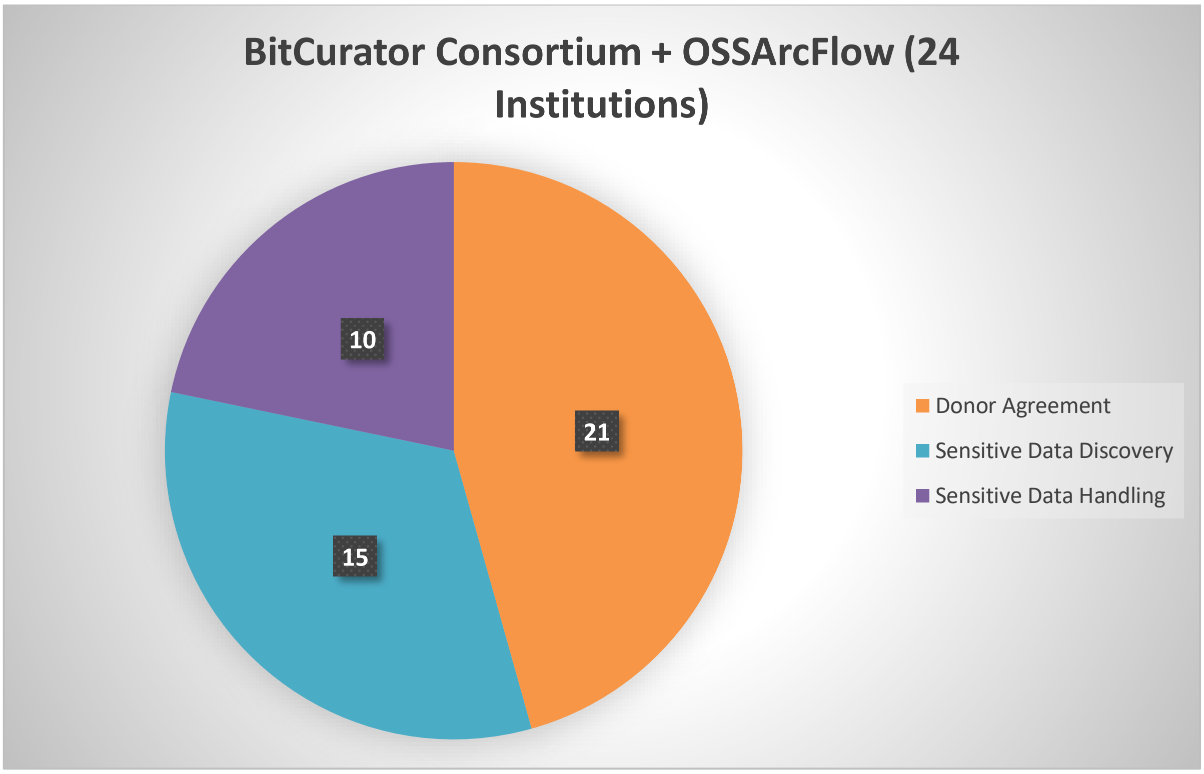


Figure 36 - Combined Criteria Breakdown (24 Workflows)

Figure 36 illustrates how each of the criteria was met across the combined set of 24 workflows between the BitCurator Consortium and OSSArcFlow. This shows that only 10 of the institutions were handling sensitive data effectively; 15 performed sensitive data discovery; 21 had an adequate donor agreement process.

### Exemplary donor agreement and decision-making

Outside of these public collections, a draft workflow was discovered, provided by Barrett, (2017) which focuses on accessioning, appraisal, and processing. Although only a draft workflow, and therefore not included in the tested datasets, it provides an in-depth view on how the institution deals with the ingest process and displays considerable detail regarding interactions with donors; there is also a distinction between solicited and unsolicited material.

From the accession of donated material, many decisions are being made that lead to definitive solutions, such as, termination of the process should the material not be selected for retention, including what happens to the material in this event.

If the donated material was provided by an unsolicited drop-off, the donor is given a questionnaire. Based on the answers provided, an archival appraisal takes place to deem if the material will be selected for retention. If the data are not desired, the donor is asked if they want the discards, if so, the material is made available to them, else, the material is disposed of. This terminates the workflow.

If the donated material was solicited, the next decision is based on if the "deed of gift" is signed.  If it is signed, it is then appraised for monetary value, and if this is to be true, the material is then taken through the core processing of the workflow. If the deed of gift is not signed, the material is packaged, accession records are created, and depending on the type of media, digital or physical, disk images may be created and then everything is sent to storage. Unsolicited material that has been approved for retention is also taken through this process, advancing to the node in the workflow that follows the deed of gift check.

All the correspondence with the donor is documented together with the stored material. The next node in the workflow checks if the deed is yet signed, terminating if it has not. If it has been signed, but the material is not a priority at the time, the workflow is terminated. If both criteria are met, the workflow proceeds as normal.

The following workflow processes take two avenues, one where forensic processing is required and one where it is not. Overall, the workflow is quite good, providing much detail and includes good handling of data and events. The depth of the accession and appraisal processing is exemplary and is something that should be considered in existing workflows or accompany them as a sub-workflow. A node within an existing workflow can point to a sub-workflow which expands on the processing involved. From this, there are two workflows, one that is high-level, only providing the basic nodes to show the overall process in a simplistic form. The other is low-level, expanding on processes where detail and transparency is needed.

## 7.3 Workflow Enhancements and Visualisations

The information gathered from the workflows aided in the development of provided solutions to the issues regarding lack of transparency and proper sensitive data discovery and handling as identified in Chapters 4 AUSTRALIAN LAW IMPLICATIONS and 6 DIGITAL FORENSICS – SENSITIVE DATA. The concept of enhancing workflows to introduce the required solutions is done in a manner that will provide multiple levels, both high and low, allowing main workflows to be expanded on with sub-workflows, reducing overall complexity. A modular design is used so the enhancements can be amended to existing workflows, adapting to existing designs. With this approach, existing workflows will not be impacted, but accompanied by expansive sub-workflows that can be pointed to or injected directly with a few changes to make them fit.

All suggestions are done so with considerations regarding institutional resources. Suggestions are made on tools that are needed to meet the requirements, such as digital forensic tools to identify sensitive data. The institution is free to select a tool of choice where there are open-source solutions that can achieve the same results as paid proprietary choices. One thing that is certain is the processing time required when using digital forensic tools, as seen in the experiments of Chapter 6 DIGITAL FORENSICS – SENSITIVE DATAis a luxury not always available. This is taken into consideration and any solutions that include forensic processing will try to avoid disrupting core processes heavily. There will of course be some need for compromise, but there will be cases where the suggested processes can be run simultaneously or at the institution's convenience. There are many variables to consider, many different outcomes, and it will be a learning experience as to what the best course of action is. This is something that will be unique to each institution.

This section focuses on the visualisation of enhancements that are presented both visually and descriptively. Elements such as notation and design are discussed and presented as options, rather than definitive solutions, allowing for flexibility.

As the targeted audience may not have a technical background, designing easy to follow diagrams was necessary. Hence the decision was made to reduce complexity and technicality as much as possible to allow flexibility in design and implementation.

The first element to be discussed is design options, such as notation, layout, and overall characteristics. The design of the workflows analysed will be evaluated and presented where good design is determined to be used as an example and baseline.

Each set of workflows that has been analysed contains different design elements. Each of these has been reviewed to determine good attributes to consider or be mindful of in the design of the enhanced workflows.

### 7.3.1 Notation and Design

The first set of notation comes from the 2012 set of the BitCurator Consortium dataset, presented in Gengenbach, (2012). This legend shows the different shapes, arrows, and forks used and the activity they each represent:

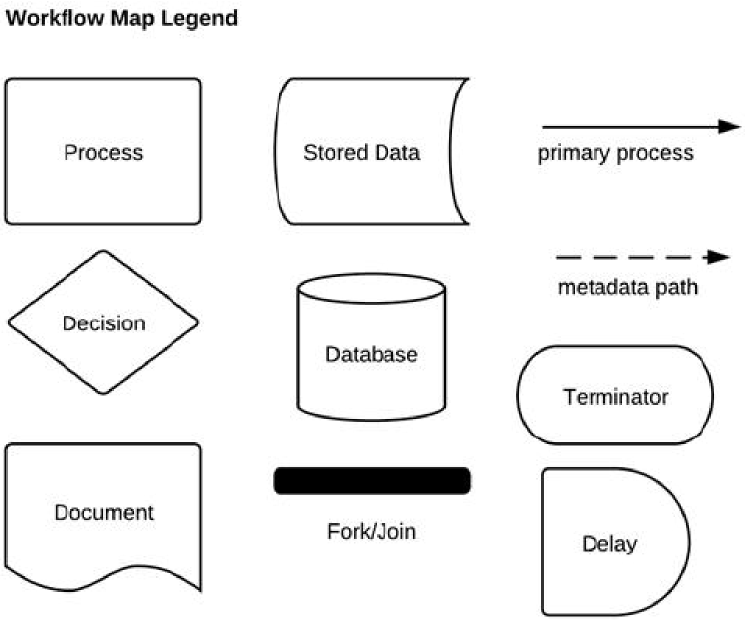


Figure 37 - Workflow Notation - BitCurator Consortium, (Gengenbach 2012)

This form of notation can be expanded further, but in recent workflows, some of this remains unused or has been condensed. There are cases where additional elements have been added, such as in the OSSArcFlow workflows which use the notation shown in Figure 38 and Figure 39

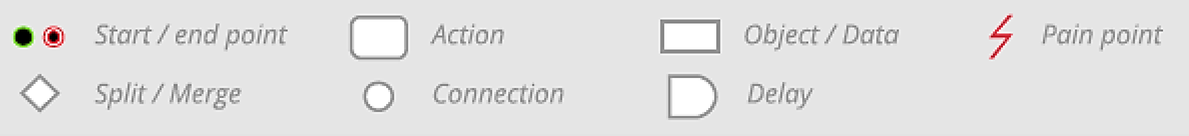


Figure 38 - OSSArcFlow Notation

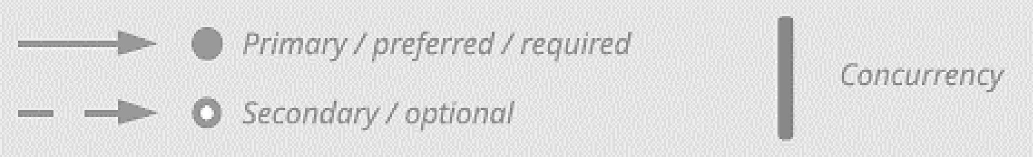


Figure 39 - OSSArcFlow Notation

The two sets of notation share similarities and contain differences, the main change being terminology. “Objects” and “Data” have been condensed in the second set of notation. The secondary arrows have been labelled more broadly to indicate any secondary or optional processes, rather than metadata specifically as shown in Figure 37 - Workflow Notation - BitCurator Consortium, (Gengenbach 2012). Although alternative terminology is used, the notation for “Fork/Join” and “Concurrency” behave the same way, this also applies to decisions which have been labelled as “Decision” and “Split/Merge”. The addition of “pain points” in OSSArcFlow notation (Figure 38 - OSSArcFlow Notation), indicated by the red lightning bolt symbol, are not used to represent a process within a workflow; they serve as a note to the reader where issues exist at any point in the workflow.

The second design to be discussed is one that implements swimlanes, which have evolved in recent workflows to alternative designs. Regardless of what design is to be used, swimlanes or any design in which users and systems can be visualised, should be considered as they provide additional information on who or what is responsible for each process within a workflow; they will also prevent messy diagrams and keep them on a linear path. However, linear design is not always optimal in complex designs with multi-level decision-making, but alternative swimlane design can help with this. Although duplicate and redundant nodes within workflows should be avoided, it may be the only option to maintain the linear flow within some complex workflows.

Swimlanes can be used vertically or horizontally, traditionally presented in vertical formats in legacy models with horizontal formats being preferred in modern design, typically read left to right. As an example, Figure 40 was created to illustrate horizontal swimlanes:

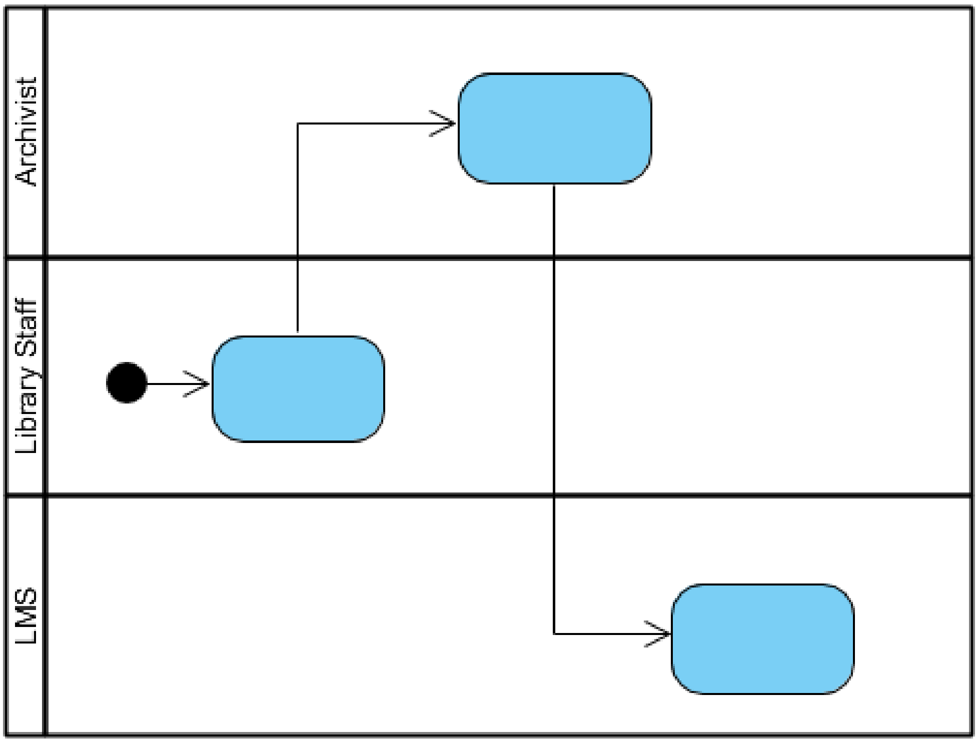


Figure 40 – Horizontal Swimlane Example

In this example, the lanes indicate the different systems and users; LMS, Library Staff, and Archivist. The processes within these lanes are handled by that lane's system or user. Processes can be forked and handled by two or more lanes at once as seen in Figure 41.

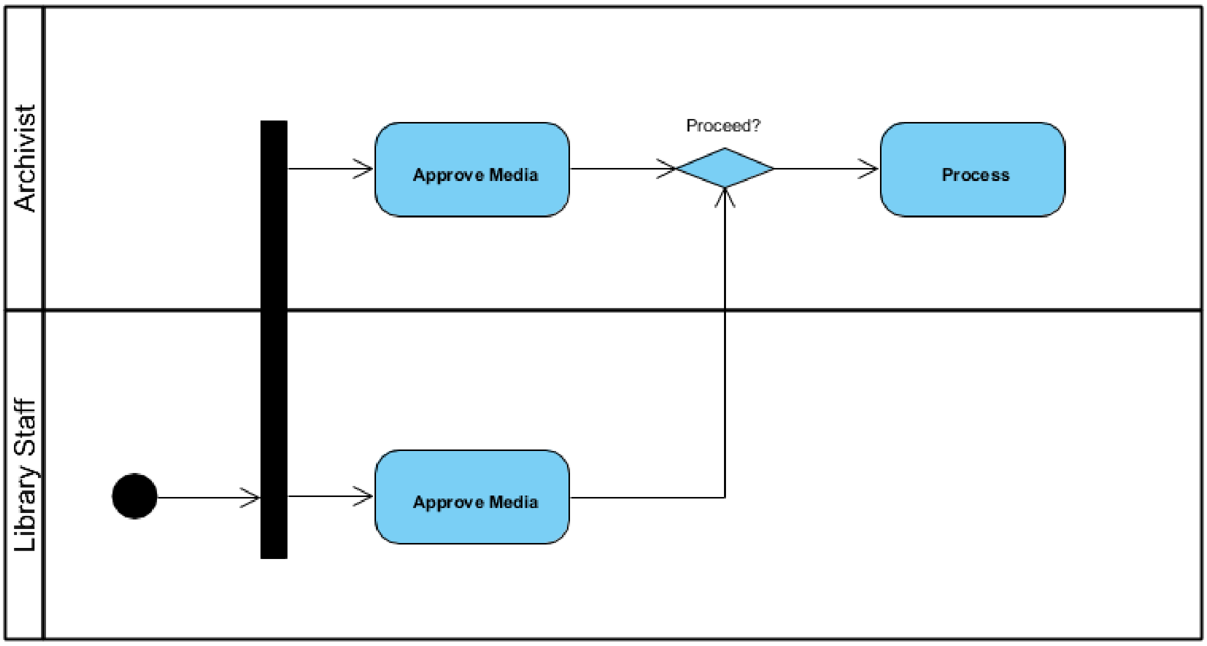


Figure 41 - Horizontal Swimlane Fork Example

In this example, the fork is indicated by the black bar that spans two lanes. An initial procedure requires both the library staff and archivist to approve the media before it can proceed. There are multiple ways in which a setup such as this can be displayed. The user may wish to fork the lines or create duplicate entries in each lane. The overall style of the diagram and the position within may dictate the best method. The choice of modelling software may also factor in the presentation of visualisations. Visual Paradigm CE 15.1 was used for these examples.

The OSSArcFlow workflows are quite extensive and use significant horizontal space, therefore, they have used an alternative design to swimlanes. This design allows the user to maintain clear understanding of each user and system as the workflow progresses, requiring the reader to scroll across dependent on the size of the monitor. With a standard swimlane design, as the reader scrolls, they lose sight of the swimlane legend, which can be hard to remember with many systems and users. The alternative design presented in the OSSArcFlow workflows prevents this issue. This design has been separated in two parts, which are presented at the top and bottom of the workflow. The top (Figure 42) indicates the software or tool used to handle the process and the bottom (Figure 43) indicates the user or staff member.

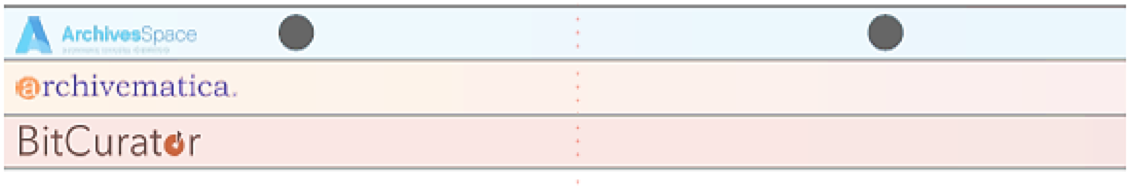


Figure 42 - Alternative Swimlane (Software/Tools)

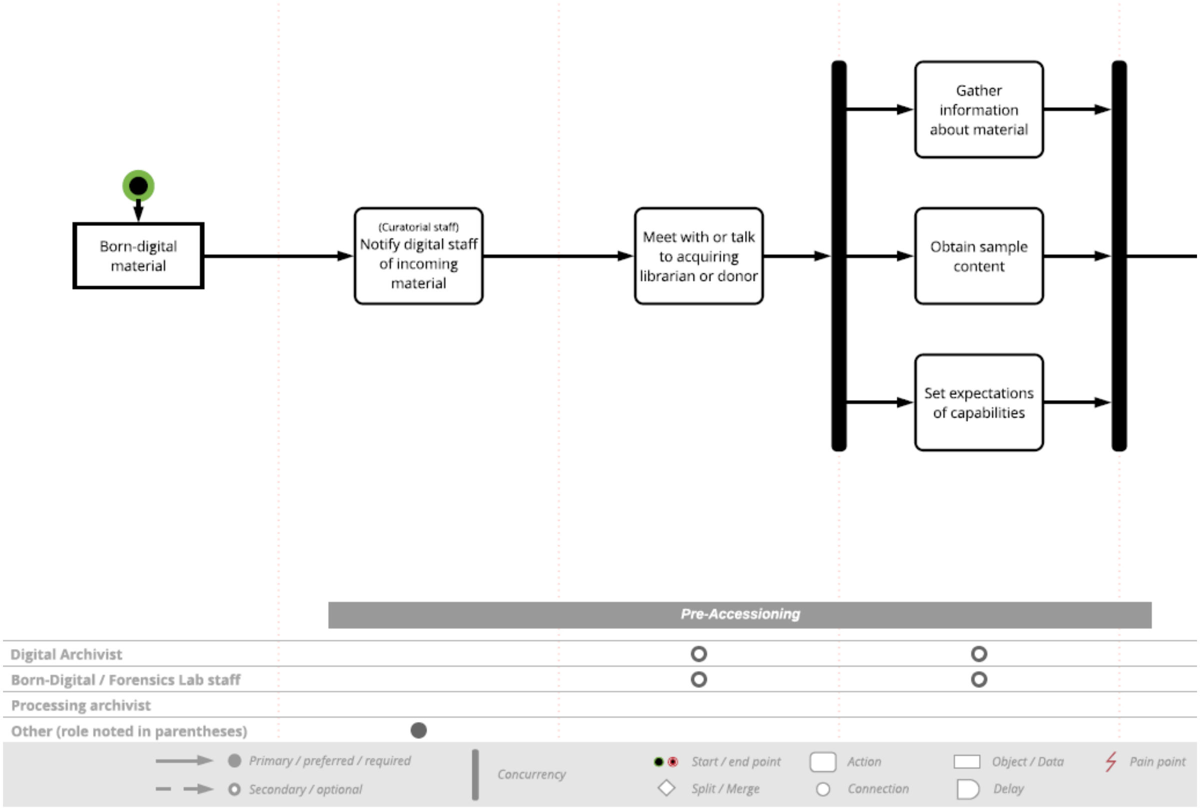

Born-digital material
(Curatorial staff)
Notify digital staff of incoming material
Meet with or talk to acquiring librarian or donor
Gather information about material
Obtain sample content
Set expectations of capabilities
Pre-Accessioning
Digital Archivist
Born-Digital / Forensics Lab staff
Processing archivist
Other (role noted in parentheses)
Primary / preferred / required
Secondary / optional
Concurrency
Start / end point
Split / Merge
Action
Connection
Object / Data
Delay
Pain point

Figure 43 - Alternative Swimlane (Users/Staff)

Solid circles are used to indicate which lane is responsible for each primary process. Secondary and optional are indicated by regular circles. The legend labels are repeated throughout the workflow, removing the need to scroll back across to check or requiring the reader to remember it. This design eliminates the need for unnecessary forking or node duplication as simultaneous users or systems can be indicated by adding a circle to multiple lanes. For example, if both the “Library staff” and “Processing Archivist” (Figure 43) were required for a particular process in the workflow, solid circles are applied to both.

Both designs have their merits and deciding which to use would be determined by the length and complexity of the workflow. For a straightforward workflow, traditional swimlanes may be adequate. If the workflow is large and complex, made up of many systems and users, the

OSSArcFlow design may be better suited. In the event of growth and expanding workflows, it is not difficult to migrate from one format to the other.

The design elements presented serve the purpose of showing potential design ideas to collection institutions. This is beneficial as some institutions may have yet to implement a dedicated workflow or may have a basic one in place which could be improved with these designs. For example, the alternate swimlanes present a different approach which will allow the institution to better visualise their process with increased efficiency and clarity. However, as the enhancements are additions and improvements to existing workflows, they are presented in a way that can be implemented at the institution's discretion.

The choice of notation also depends on what best suits each institution. Whilst various elements exist for describing different actions or objects, such as different lines for various objects and data, they are often unused, and the diagrams are simplified. Simplification is optimal as the workflows for digital preservation describe *what* is being done more so than *how* it is being done. Workflows are a guide through the process of which should be followed, but there are times where there is a need for flexibility. Therefore, the notation discussed is there to provide options to be considered when implementing workflow enhancements and potentially re-designing existing workflows to accommodate said enhancements.

### 7.3.2 Enhancements

The three core areas of focus are:

- Donor agreements
- Sensitive data discovery
- Sensitive data handling

The sensitive data enhancements are presented as standalone additions with consideration to the existing digital preservation workflows. Whilst the goal is not to impact existing workflow procedures, there are instances where improvements are needed within the core processes that will also benefit the enhancements. Donor agreement enhancements are the most likely to alter existing procedures as they are the most prominent processes amongst the institutions.

Minimal interruption and impact on the core processes are of the utmost importance when considering the suggested enhancements.

### Donor Agreements

Donor agreements are the area in which the majority of institutions investigated had approached effectively, as seen in the workflows analysed in Chapter 7 WORKFLOWSand discussed in Section 5.2 Australian Institutions. Therefore, the following solution involves visualising what already exists in current procedures and ensuring the donor process is amended to accommodate sensitive data discovery and handling.

Improvements must start at the beginning of the workflow and are done by enhancing the processing and decision-making regarding donor communication and agreements. With a thorough process to gather information about the donor and their material, the need to revisit this process in the later stages of the workflow can be avoided. This is done by ensuring that as much information is acquired about donated material where possible and establishing the risk level associated with the data potentially residing in the material.

Establishing ownership and any "what if?" scenarios is equally as important, allowing some issues to be handled in advanced should they surface. This mainly involves ensuring that the approach to be taken in the event of unforeseen discoveries or if contact with the donor cannot be established is well documented.

The visualisation in Figure 44 shows a basic flow of process in the event that material is donated to the institution through the proper channels and directly by the donor, indicating there is a documented process such as completing an online form or requesting an appointment. There are instances where collection institutions receive donations through unconventional means and have unique procedures for handling such events. These instances are important as there are uncertainties with the material and potentially the donor, which can go undocumented and forgo the correct processing procedures. These events are likely isolated to smaller institutions, such as digital archaeology labs within a university, where drop-off donations occur.

The donor agreement workflow designed in Figure 44 is a mix between high and low-level, with room to expand on certain nodes. The important part of the diagram is the interview process and ensuring the right information is gathered. This involves additional participation and involvement from the donor and staff member conducting the interview.

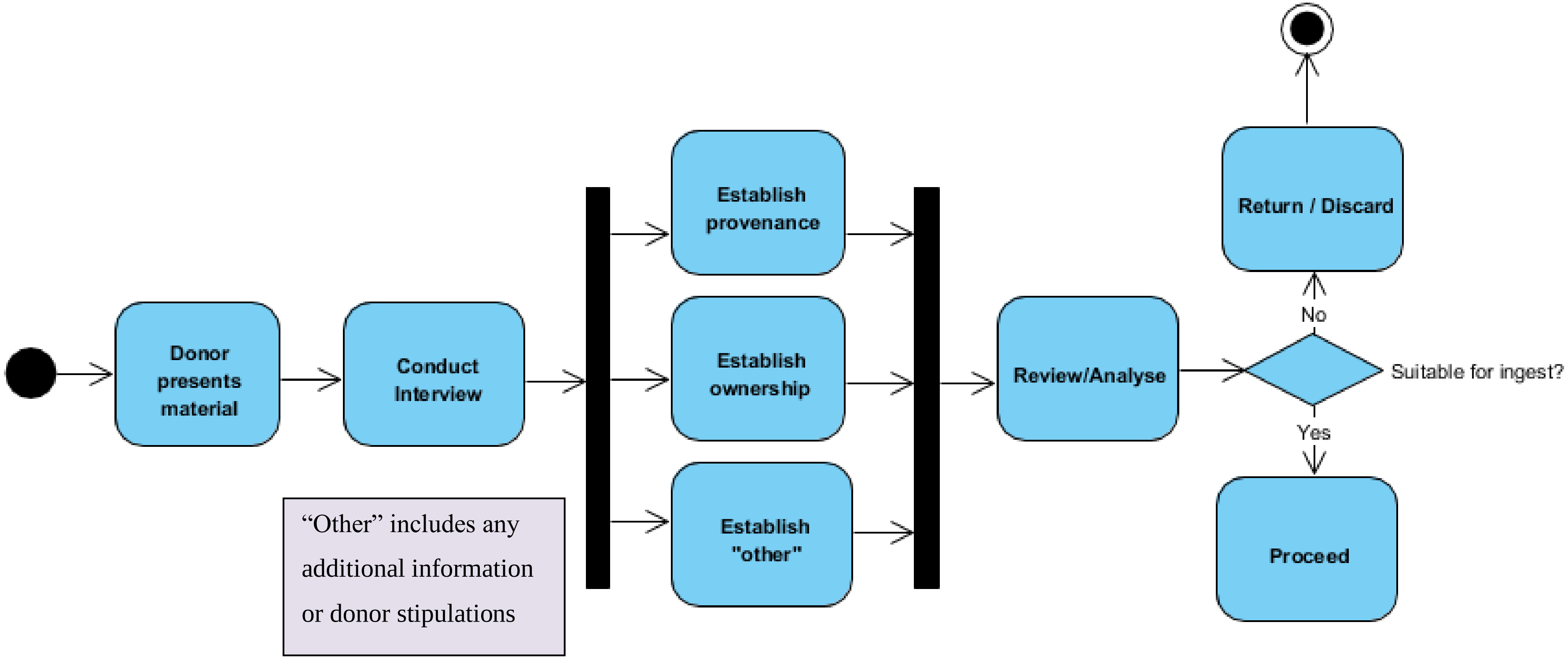


Figure 44 - Donor Agreement

The (Conduct Interview) node shows that the interview process involves establishing provenance, ownership, and any additional information or donor specified stipulations (Other). However, this process may not always be possible in circumstances where the donor is no longer available or does not wish to be interviewed. Depending on how the donor came to be in possession of the material they are donating, they may not know any detailed information regarding its provenance.

Once the interview is conducted, all the information gathered or missing needs to be reviewed and analysed to determine the worth to the collection and establish a risk assessment. This is visualised with the two forks, as the information from the interview is used to establish the three criteria (provenance, ownership, other), which are then forked into a review and analysis process.

The essential information needed includes how the donated material relates to the donor, how the donor acquired it, and if possible, the chain of custody. It is also important to identify how the donor handled the materials throughout their ownership as this may provide clues about any potential metadata creation and changes.

By establishing provenance, it will give insight into any potential sensitivities that may be uncovered. This will also determine how sensitive information is to be handled if it is discovered and whether any further interaction with the donor is required.

Within the workflow diagrams created, certain nodes such as “Proceed” indicate the workflow continues through to the next steps of standard processing. The diagrams presented are subsections of the workflow and expansions on existing nodes. The “Return/Discard” node is a high-level representation of a process that may involve many steps and decision-making. These processes are presented in this manner in efforts to not change or suggest change to existing core procedures.

It is known that not all donations can be handled within the best-case scenario. Some institutions may receive unsolicited donations or drop-offs and these situations need to be handled with extra care and thoroughness. Figure 45 presents the suggested method of handling unsolicited donations:

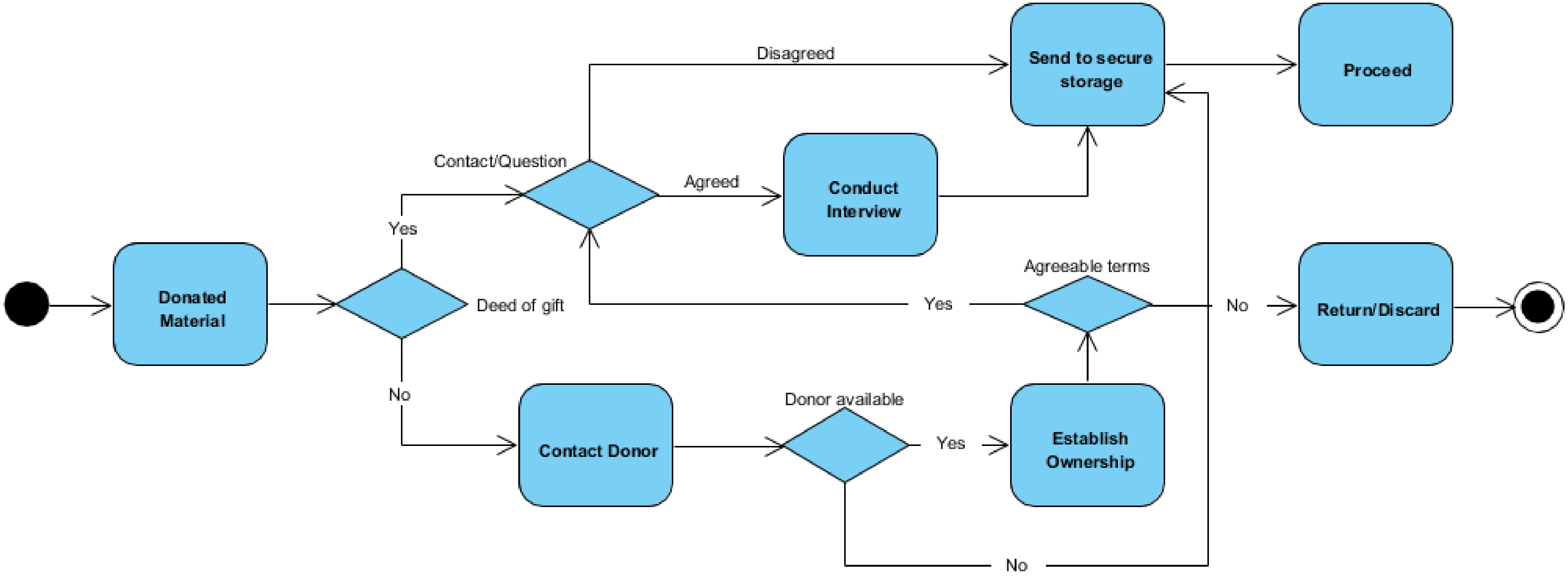


Figure 45 - Unsolicited Donation

Multiple pathways exist in Figure 45 due to the multi-level decision-making that exists to ensure all contingencies are considered. The pathway chosen depends on whether a deed of gift is signed, which is identified through the first decision node following the “Donated Material” node. With the deed signed, this transfers ownership from the donor to the institution. This presents the pathway located at the top of the diagram.

On this pathway, a second decision node is presented (Contact/Question). This involves contacting the donor to gather any further information via an interview. It is assumed the donor is available as they have signed the deed and provided their details. If contact is successful, any concerns regarding ownership should be clarified. Once agreeable terms have been met, usually when the donor passes on ownership to the institution and no longer holds any rights, or at least has reasonable stipulations, the donor should then be presented with the option to participate in an interview. If the donor agrees to participate, the pathway continues to the “Conduct Interview” node which is a high-level representation of the interview process from Figure 44. This includes establishing provenance, ownership, any additional information, or donor specified stipulations (Other).

Once the interview has been concluded with satisfactory results, the material is sent to secure storage along with any accompanying data derived from contact with the donor.

If the donor disagrees to the interview or cannot be contacted, the material moves to secure storage, bypassing the “Conduct Interview” node. This is an acceptable option as the deed of gift has been signed in this instance. The interview process is there to gather any additional information and clarification but is not required to proceed through the workflow.

If the deed of gift was not signed, the bottom path is taken. The next step is to contact the donor. If the donor is not available or does not wish to participate, the donated material is sent to secure storage before a decision can be made. If the material is to be added to the collection or removed, the workflow follows the institutions’ established procedures accordingly (Proceed).

If the donor can be contacted, ownership must be established. If agreeable terms can be met, indicated by the decision node following the “Establish Ownership” node, the workflow then moves to the top pathway and proceeds as if the deed has been signed. If terms cannot be agreed upon, the material is returned or discarded based on the donor’s request.

All outcomes lead to the secure storage or the discarding or return of the material to donor. Secure storage checkpoints exist to ensure no material is processed without proper decision-making, analysis, and consultation with legal if required. Certain pathways throughout this workflow collect more information than others, meaning once the material is sent to secure storage, the time it spends there before proceeding into processing is based on the information collected. The workflow then proceeds from this step, approaching the stage involving the enhanced suggestions of sensitive data discovery and handling.

### Sensitive Data Discovery and Handling

The workflow enhancements presented in this section aim to improve the discovery and handling of sensitive data. This is achieved by incorporating digital forensic tools and methods into existing workflows.

The designs of these enhancements are extensive and have therefore been broken up into sub-diagrams where necessary. Some of the nodes within the core workflow are represented as sub-diagrams because they are not core features of the digital forensic process but are still required. For example, the first node present in the sensitive data discovery (SDD) workflow (Figure 47) is “Initial Forensics”. This node is represented in more detail in its sub-diagram (Figure 46) that expands on the initial forensic processing as it is a fundamental step.

Data is broken up into two categories: “Target” and “Other”. This differentiates the data the institution is actively seeking on the source material and any other data that may reside on the media. This relates more so to donated material that may be delivered on hard drives and any other storage media containing additional files. This assumes enough information was provided with the donation to indicate what may be of interest to the institution, establishing the “Target”.

In the “Initial Forensics” workflow (Figure 46), there are two paths. The first follows the standard media procedure of preventing changes being made to the original source and creating a way to check if changes do somehow occur. This process occurs after the donor agreement has taken place and is the first point of ingest.

The second path is there to indicate the option of alternative methods when dealing with unique and complicated media requiring deviation from standard procedure. The process that takes place through this pathway may be unique each time, but it is assumed that the same precautions are taken where possible to ensure no changes can be made and that checksums are performed to verify this. This visually represents the need for flexibility and acknowledges how each case may be different, requiring alternative methods. The standard procedure should be met whenever possible, that is the creation of an image with file verification (checksum), ensuring the original data remains in its original state.

Flexibility, choice, and thorough decision-making are strongly emphasised in all suggested workflow design. As this is not a core part of the SDD, this suggested workflow can be adapted and changed at the institution’s discretion. It is provided as a baseline and to promote transparency in that all processes should be visualised or described.

The main workflow for SDD (Figure 47) is quite extensive and can be represented in various levels. In its current state, it is made up of multiple levels, but it is also flexible in that some areas can be reduced to higher levels and some can be expanded on.

Another node within the core of the SDD workflow that is expanded on in a sub-diagram (Figure 49) is the “Evaluate” node. Most of the paths within the workflow must pass through this node, therefore, it is expanded on in a sub-diagram rather than adding its complexities to existing workflows, promoting efficient design. To have this node described in full within the workflow would increase its complexity and size. Splitting the workflow up into smaller diagrams should not be an issue when providing access to this information if the sub-diagrams are presented with the main workflow and are not stored in obscure locations requiring the user to navigate several web pages to access them.

The main features of the SDD workflow are as follows:

- Multiple pathways for data with different processing requirements
- Risk analysis and extensive decision-making
- Sensitive data discovery
- Sensitive data handling
- Secure storage checkpoints
- Donor incorporation
- No gaps where information can slip through without evaluation
- Ensures thorough investigation of source media

The SDD workflow contains a pathway where if there is no risk involved and the target material is known, it is allowed to bypass the bulk of the decision-making and evaluation. The process may then proceed to the later stages of the workflow. This will rarely occur and only in cases where the material being processed is basic, straight forward, and cannot contain any invasive metadata.

On the same pathway, if the decision is made that risk may be present, the material is sent through to secure storage where it will wait for the “Other” material to be processed. In this case, the risk is that the “Other” material may reveal information about the “Target” material and no further processing should occur until all the data have been analysed. Examples have been provided throughout this study where sensitive information can have an impact. From this, the material must pass through the “Evaluation” node. This ensures the risk is assessed properly before any access is provided.

Both the target and other data are initialised by a SDD node. This is where some options are presented to the institution. Some form of forensic processing needs to take place here, making use of software such as those that have been reviewed, or alternatives. There will be some cases where a full investigation making use of a forensic package such as Autopsy is not required. If the material being processed is strictly text based, then the features of Autopsy would not be required. Using a tool such as Bulk_extractor is better suited for such material as it will scan the contents of the text and any metadata within the files.

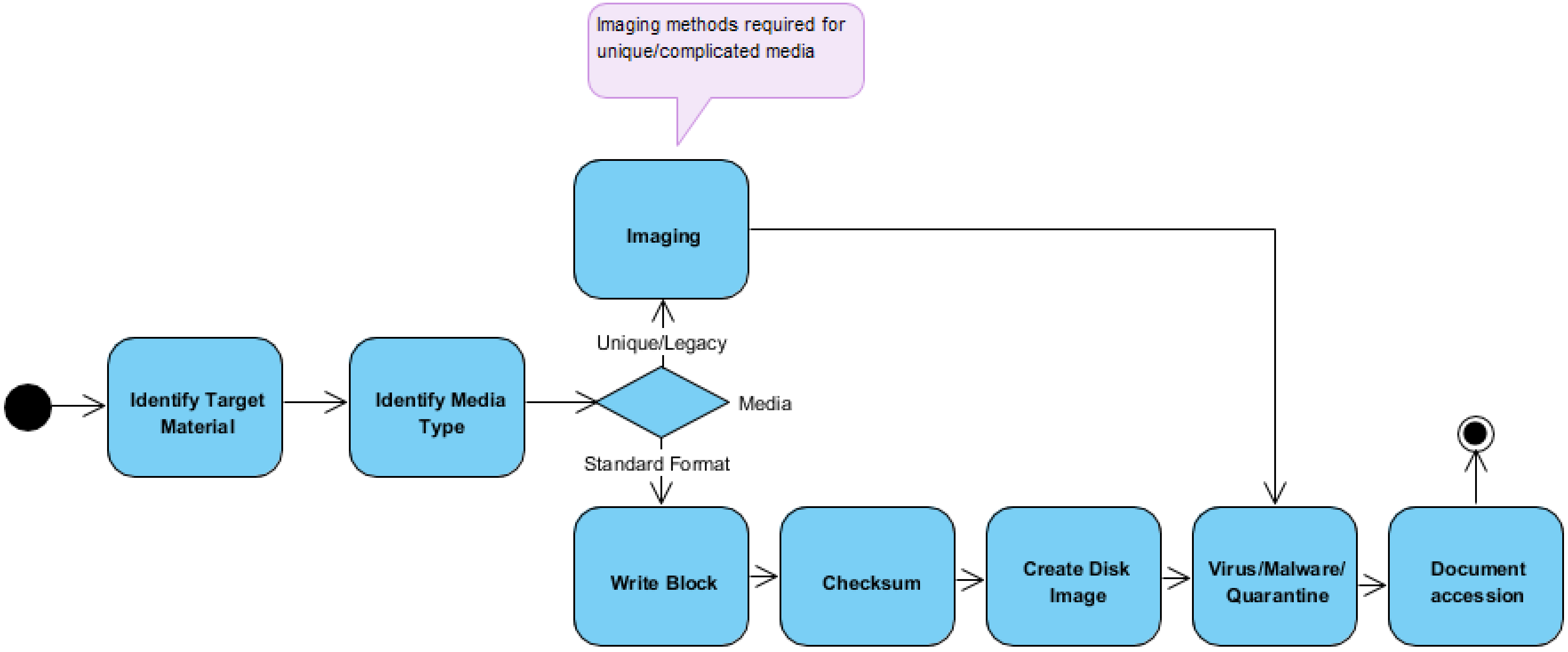


Figure 46 - Initial Forensics Sub-Diagram

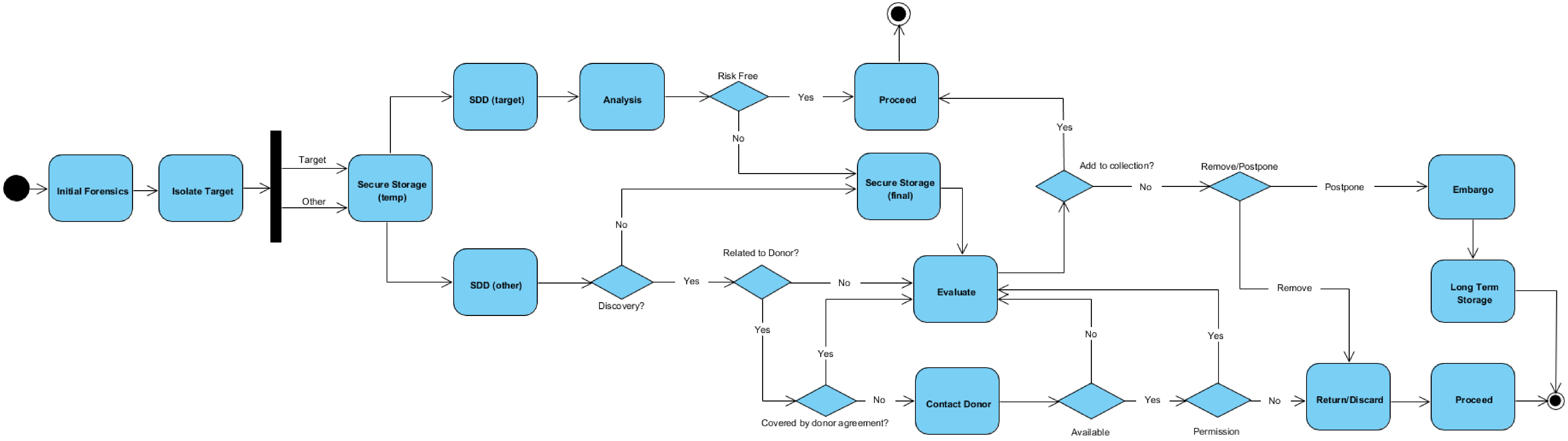


Figure 47 - Sensitive Data Discovery and Handling

At a minimum, sensitive data revealing personal information about the donor or the target of interest must be top priority as these are the most likely types of data to cause issue. If there is a risk to the authenticity of what is being published, the investigation should aim to determine this by establishing provenance metadata, especially change history and ownership data. In any case, the “Other” material should be processed in the event that something relevant is discovered. The processing of “Other” secondary material can be run concurrently to the “Target” material, or at a more convenient time where more resources are available. However, delaying the processing of “Other” secondary material should only be done with complete confidence that the target material can be processed safely without risk. If emulation is required, the digital forensic tools mentioned will allow a better understanding of the system and the environment the source material came from.

Regarding the remaining paths of the workflow, there are three ways in which the process may continue or terminate. The first is the exit for the “Target” material if there are no risks involved. The second pathway, depending on the attributes and processing required, may exit through the first exit or the second exit. On this pathway, if the “Other” material is processed and there is no SDD, it is sent straight to secure storage. This material will be evaluated for its worth to the collection, or it may in fact be related to the “Target” material, in which it will then be processed accordingly.

If there is no SDD, there is no need to proceed to the node within the workflow that checks if the data has any relation to the donor. It is important to differentiate these types of SDD as the impact and the procedures needed will vary based on how the SDD relates to the donor or “target”. In some cases, the donor may in fact be the “target”. For example, an artist may wish to preserve their work within a collection. There are many variables to consider, but no risks should be taken to publicise information unless it benefits the collection, and any impact or exposure is assessed.

If there is a SDD that is not related to the donor, it is sent to evaluation immediately. If related, checks are in place to consult the donor agreement and make contact if the agreement does not cover the content in question. If contact cannot be achieved, the content is sent for evaluation. If the donor can be reached and gives their permission to use the discovered data, the content will proceed to evaluation. If permission cannot be granted, the material will not proceed and must be returned or discarded based on the donor’s decision. This pathway to the second exit is also followed for material that has been evaluated and an embargo or a waiting period for whatever purpose is set, resulting in the content being moved to long-term storage.

Before the “Evaluate” workflow is discussed, an example of how the secure storage may work is shown in Figure 48 as it would need to be included should full transparency be achieved.

Whilst the Secure Storage workflow (Figure 48) diagram is simple, it does have a critical step. The two pathways that flow through this are that of the “target” material with assumed risk, and the “other” material that has no sensitive data. Note that this is only for the “secure storage (final)” and not the temporary secure storage seen at the start of the SDD workflow. The purpose of the temporary storage is for items to remain in a location that is isolated from other systems, data, and restricted from unauthorised users.

The critical component of this sub-workflow is that target material is halted until “other” is processed. The reason behind this is “target” material should not proceed in case something within the “other” material is related to, or may influence the “target”, which may lead to changes or realisations that need to be addressed before proceeding. It would be unwise to publish the “target” without investigating the material which accompanied it on the media on which it was delivered. The halt will remain active until it is deemed safe to proceed to evaluation.

Any “target” material deemed risk free, meaning there is no way any data discovered in “other” could alter the “target” material in any way, bypasses this process and proceeds through the workflow.

Institutions may enforce another secure storage block later in the workflow before access is granted as a precaution to ensure all checks have been met and evaluations have been performed adequately. There is no harm in taking extra care; however, there may of course be resource limitations that hinder this regarding the movement of large amounts of data between storage locations of which can be time consuming. Performing taxing I/O tasks will reduce system performance whilst the transmission of data takes place.

The institution is of course in control of how much detail they wish to divulge regarding what happens in their secure storage. They may in fact already have policies in place to which they can refer, stating that any data sitting in secure storage will follow these policies. It is of course suggested that this be visualised, especially if the institution has a unique and novel way of handling secure storage as this would be greatly beneficial for learning institutions.

Following the secure storage workflow is the “Evaluate” sub-workflow diagram (Figure 49).

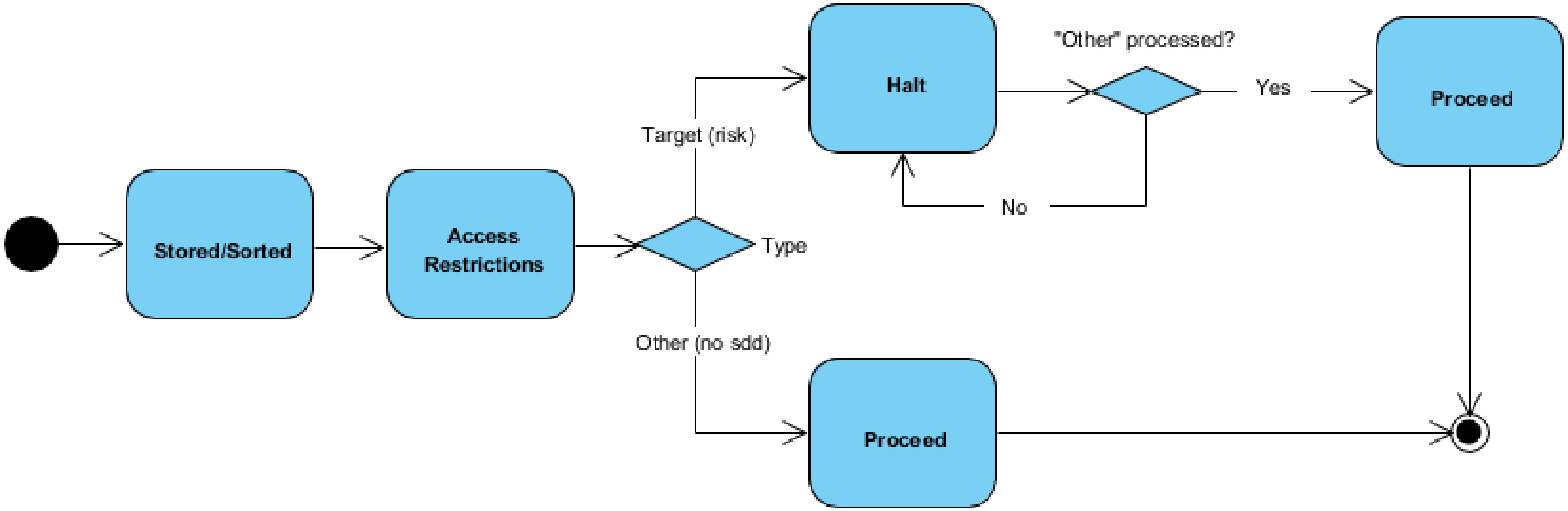


Figure 48 - Secure Storage (Final)

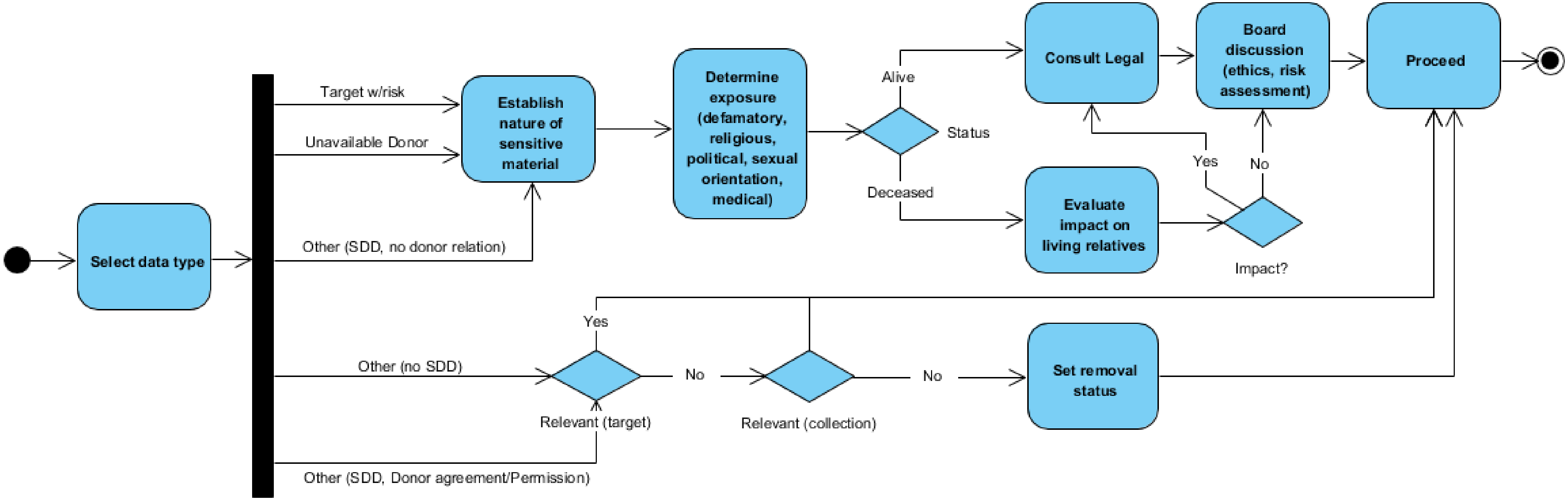

Select data type
Target w/risk
Unavailable Donor
Other (SDD, no donor relation)
Establish nature of sensitive material
Determine exposure (defamatory, religious, political, sexual orientation, medical)
Alive
Status
Deceased
Consult Legal
Board discussion (ethics, risk assessment)
Proceed
Evaluate impact on living relatives
Yes
No
Impact?
Yes
Other (no SDD)
No
No
Relevant (target)
Relevant (collection)
Set removal status
Other (SDD, Donor agreement/Permission)

Figure 49 - Evaluate - Sub-diagram

The purpose of this workflow is somewhat of a failsafe. Each item in secure storage or moving through the workflow must come through this node and have its relevance determined. Most data will pass through the entire evaluation process; however, there are two instances where this is bypassed. If an item from “Other” had no SDD, then it is checked for relevance against the “Target” and collection, if there is no relevance or value, removal status is set, and the discard procedures are handled once the evaluate workflow proceeds. This applies to “Other” material that has SDD, but the donor agreement definitively allows it or the donor themselves have given their permission upon contact, in which case, if it is relevant to the target or collection, it may proceed.

Data that may reveal defamatory, religious, political, sexual orientation, and medical information, are all candidates for potential risk and harm. If material is based on a person or group, then their vital status, and whether they are alive or deceased, are factors that must be considered. As discussed in Chapter 4 AUSTRALIAN LAW IMPLICATIONS, a deceased person cannot claim defamation or have someone claim it on their behalf. However, an evaluation must take place to determine if any living relatives can be impacted should this information be made accessible to the public. For example, medical history may reveal hereditary medical defects that could potentially reveal information on the living relatives.

Legal advice must be consulted if the individual or group the sensitive data are related to are still present or if any living relatives may be impacted by the information. If the information is not related to a person or group, but an entity such as an organisation, then the process would follow the “alive” pathway, consulting the legal department.

If there are no legal issues, the information must still be discussed and assessed from an ethical perspective. Although there may be no legal risk, it does not eliminate the fact that it may be unethical to proceed. If the information serves no purpose and is somewhat defamatory, it is preferable to keep this information from public access. If the information serves the purpose of the public’s best interest, this then becomes a grey area that needs to be discussed by a board or committee made up of individuals across different expertise and disciplines. Factors such as the viability of preserving the data in question, potential future legality and ethical concerns, the value of adding said data to the collection, and other factors that may determine the course of action to take.

It may also be in the institution’s best interest to investigate the change history of relevant laws and how they have adapted over time, giving precedence to any potential future changes. What may be an ethical issue now could become a legal issue in the future. Therefore, it is emphasised that although exemptions exist for the collections of these institutions, the laws that surround them may be suitable guidelines. If the relevant laws are considered and followed, the institution will be protected from future changes where their exemption may be jeopardised.

### 7.3.3 Summary

In the sections on workflow evaluation and enhancements, workflows from institutions outside of Australia were reviewed along with the data collected from the questionnaire results returned from the participating Australian institutions. The workflows were reviewed based on their design, if and how they handled sensitive data discovery, and how much of their process was visualised.

This review process gave insight into examples of both good and bad design, but it also allowed for comparisons to be made which gave some indication as to why there are instances of bad design. Resource limitation, data intake, and the need for preservation, are all factors that may influence how an institution is performing. The data provided from the Australian institutions supports this statement as the results came from different ends of the spectrum, that is, institutions that are performing at a higher level of preservation to those that are in their infancy.

Whilst it is clear digital forensics does not have a strong presence in Australia, it is growing in most of the institutions. Some of the institutions, at the time of participating, were starting to adopt and incorporate digital forensic tools and methods into their preservation workflow. Others are still shaping their preservation strategies, but it is certain that the need for preservation will increase, so too will the need to adopt new strategies.

Therefore, although it has been re-iterated throughout this study, it must be stressed again that regardless of whether workflows are an accurate representation of the institutions process, they are often publicly available. These may be in online articles to be used as guides, or on an institutions website under their public documentation where users can view information on their policies, procedures, and workflow. If they are accessible and are not accurately representing the institution, it may lead to incorrect assumptions by those wishing to add their material to the collection or for peer institutions that require an example to follow.

Some institutions may not benefit from the suggested enhancements as their intake levels and preservation needs are still quite low in comparison to some of the larger institutions. Institutions such as these are likely to learn from the other institutions that are leading in the field, which transparency can help determine. Therefore, it is important for the mature institutions to consider adopting these enhancements, so when the time comes for other institutions to progress, they will have a better example to follow, resulting in better implementation.

Therefore, transparency is stressed, it not only helps other institutions, but it may also help donors decide where they want their material to be preserved based on how the institution handles its collection. If someone is to donate something of relevance to the community that may also be personal in nature, they may wish to know exactly how their data are to be handled. If this is public information, it makes the choice in institution easier for the donor. If the donor is presented with a choice between two institutions, one that is transparent in their inner workings, and one that is not, if the transparent institution meets the needs of the donor, there is no reason to consider the other.

Regarding workflows that are accurately representing their institutions but are missing critical components to achieve adequate sensitive data discovery and handling, it is crucial for them to have an appropriate example from their peer institutions.

The enhancements presented in this chapter aim to provide a starting point for institutions that need to implement or improve their digital forensic processes. The main enhancement is the sensitive data discovery and handling workflow, which aims to provide the following:

- Multiple pathways for data with different processing requirements
- Risk analysis and extensive decision-making
- Sensitive data discovery
- Sensitive data handling
- Secure storage stops
- Donor incorporation
- Elimination of gaps in the evaluation of information
- Ensures thorough investigation of source media

Sub-diagrams for specific workflow processes have been presented to reduce complexity in the core workflow and to expand on specific areas. These include donor agreement handling, unsolicited donations, secure storage, and evaluation (legal and ethical issues).

The goal is to make workflows extensive enough to cover all aspects of the preservation process whilst not allowing any gaps where data may be processed without thorough investigation and evaluation. Having the diagrams broken down into smaller sub-diagrams allows this to be achieved whilst keeping the complexity to a manageable level, making it easier for staff and users to understand. Even trivial aspects of the workflow should be visualised for both transparency reasons and because by seeing a visual representation of something, it is easier to see flaws and potential improvements. This was evident when designing the workflows as they have gone through many iterations based on this principle of studying them and seeing where improvements could be made.

All suggestions made are intended to allow flexibility and change on behalf of the institutions for which the workflows are designed. The hardware and software used to achieve the suggested enhancements is a decision for the institution. The solutions that have been mentioned in this study should be considered. There are alternative solutions which can be explored. Therefore, the emphasis is on “what” to do, the “how” is ultimately up the institution.

Aside from resource limitations being one of the main reasons for the flexible and modular approach to the suggested enhancements, another is how the workflows are seen and used across different disciplines. The very nature of digital preservation separates it from most business processes. The idea of automation, whilst beneficial, is not something that is likely to occur for some time as the human element is still essential. Systems and tools can automate the ingest process, the data extraction, access, etc., but there should always be a human analysing the output for false positives, errors, and other anomalies that tools may not detect.

There will be cases where decision-making must occur, based on ethical and moral views, and not quantitatively calculated logic, which is where the human element exceeds the capabilities of any computer algorithm likely to be employed within collection institutions. Thus, it may not be in the best interest of collection institutions to follow standards based on business process modelling nor the use of workflow management systems. Workflows should serve as flexible guidelines that suggest what to do, but not necessarily confined by the constraints of the workflow. There may be unique media in which the normal workflow operations are not applicable but are still guides as to what needs to be done, even if how it is done deviates from the workflow. Born-digital data will often present unique cases which require deviation from workflows, or new workflows entirely.

# 8 DISCUSSION

The following discussion involves a use case based on the information of each chapter to provide insight into the outcomes of this study and how they apply to collection institutions. This provides an overview of how to implement and perform the suggested enhancements and why each element is relevant. This use case also explores the options that can be implemented by the institutions. The implications and challenges, both present and future, are re-iterated and emphasised.

Every institution and their requirements may be unique, therefore, for the purposes of this discussion, assumptions are made to provide a well-balanced example.

The first assumption is based on the level of digital preservation being performed by the collection institution. An adequate level of preservation is assumed, meaning there is a regular intake of donated and curated material, as well as a dedicated preservation workflow which includes dedicated preservation tools. This ensures appropriate measures are in place for proper preservation protocols from ingest to storage. It is important to remember that when referring to digital preservation, it is meant to reflect the whole life cycle and any actions that occur during the process. This includes discussions with the donor right through to ongoing maintenance and providing meaningful access once published within a collection.

The remaining assumptions are based on what is missing and what can be improved. It is assumed there is no dedicated digital forensics being performed outside of disk image creation and the use of write blockers.

Resource restrictions are assumed to be time and money based. Funding for staff training and software is considered as is the increase in processing time with the addition of digital forensic methods.

For digital forensics to be properly implemented, some changes to the infrastructure are required. This may be done by introducing new hardware or re-purposing existing hardware. The main implementation required is a means to store sensitive data securely. This is referred to as secure storage. The purpose of this is a temporary location for data to sit idle before they are processed or are awaiting confirmation on other processed material. The depth of this is ultimately up to the institution, but it is recommended that at least two secure storage transitions are in place, with the addition of long-term secure storage for cases where embargos are in place.

This will introduce additional processing time in the overall workflow, but it is necessary. Access limitations will need to be in place to ensure the secure storage is only accessible by authorised personnel only.

With the additional storage transitions and the use of digital forensic tools, better hardware may be necessary to reduce overall processing and transfer time. Depending on the level of digital forensics being performed and the size of the data being processed, significant increases in processing time will occur.

The first part of the workflow should be unaffected by the digital forensic enhancements and will follow standard procedures. The interactions with the donors should reflect the suggestions made in this study. The goal is to capture all vital information about the donated material. The criteria for this are as follows:

- How did the donor acquire the material (provenance)?
- How does the material relate to the donor?
- How was the material handled by the donor?
- Are there ownership stipulations?
- What are the access conditions?
- What is the protocol if sensitive data are discovered?
- Is there any additional general knowledge known about the material?

These criteria provide the means to establish an estimation on the potential sensitive data that may reside within the donated material. Knowing where and how the material was acquired, as well as how it relates to the donor, will give some insight into the types of sensitive data and the implications that come with it. For example, if the donated material was found or purchased outside of the donor’s residence (or any family members), then any sensitive data discovered should not be related to the donor. If the material was handled by the donor using their personal computer, then there is a possibility that sensitive data may be discovered that does relate to them. In this specific case, nothing should be discovered that is not coverable by the donor agreement.

If the donated material belonged to the donor or a descendant, the complexity of the agreement will increase and there may be discoveries that are not covered initially, resulting in additional communication with the donor. This may lead to further issues, should the donor or next of kin no longer be available. Therefore, the donor process is critical, and gathering as much information as possible is necessary. The instances where complexities arise will be discussed further in the use case. Research questions three and four are increasingly important when no further correspondence can be made regarding donated material. Having considered legal and ethical implications, and perhaps created policies for these instances, a solution may be achieved quicker and with reduced risk.

Now that the donor agreement process has been conducted accurately, the material is then processed as the initial stages of the workflow suggest. Whatever means are in place to create an image and ensure no changes have been made to the source are assumed to be performed adequately.

At this point in the workflow the major enhancements take place. Data are differentiated in two forms, “Target” and “Other”. After the donor agreement process, the institution knows what it wants in its collection and what it is getting from the material; this is the “Target” data. The “Other” data are anything that remains on the source media that is not the “Target”. Both data are stored in the secure storage where they are readied for processing.

The workflows provided offer multiple pathways for both data types to follow with extensive decision-making to guide them. Every case will be unique in some way, but for this use case, two comparative examples will be used. The first is if the “Target” is known and classified as risk free, meaning nothing residing in the “Other” data can impact it in any way. The second is if data from “Other” may influence change in the “Target” data.

Both examples provide different options in how to handle the processing. In the first example, since the “Target” is known and risk free, this can be processed normally without having to wait for the “Other” data to be processed. This allows the institution to process the “Other” data when they see fit. Of course, these data could be ignored and returned or discarded; however it is strongly recommended that this does not happen. The implications of this have been identified in the output of the Bulk_extractor and Autopsy experiments, displaying the potential sensitivity of data, of which the risks have been explored in the research presented in Section 2.1.1 Ethics, Privacy, and Legal and Chapter 4 AUSTRALIAN LAW IMPLICATIONS. If time permits it, the processing of “Other” should be run concurrently with the “Target” data.

Note that in this case, risk free means there is no risk to the “Target” from the “Other” material. The “Target” is still scanned for sensitive data as there may be information in need of redaction, or the nature of the sensitive data may alter the destination of the material within the collection. For example, if the data indicates some form of indigenous relevance, it may not be suitable for the public collection and must be addressed. The same may occur for material of a highly sensitive nature, which may only be presented upon request within isolated and secure reading rooms.

The second example involves the "Target" data being halted within the second secure storage until the "Other" data are processed. This ensures all the data on the source media are analysed before the "Target" can proceed through to the final stages of the preservation workflow. With this approach, the risk of publishing incorrect or out of context information is reduced. Accuracy, authenticity, and completeness are standards a collection institution must uphold and being thorough is the only way to guarantee this.

Any data that does not fall under the "Target" and "risk free" categories, i.e. "Other", will be processed through the "evaluation" core of the sensitive data discovery workflow. This involves legal and ethical considerations and decision-making to ensure a proper risk assessment. As this study has shown, exemptions from legal considerations are not a guarantee in all cases. There are grey areas in which cases can be made for lawsuits such as defamation. When handling sensitive data, they must be analysed for any impact they may have on any related parties. Whilst the deceased may not be able to fight for themselves, and their descendants cannot not fight on their behalf, if the data reveal any information about the living descendants, there may be enough for a viable claim on their own behalf.

Where there are no legal issues, there may be ethical issues that must be addressed. There may be cases where the public's interest is considered above ethical concerns, and there are times when the ethical concerns outweigh the need to publish this information. This research has established that there is no standard approach to the consideration of legal implications (Question Three) or procedures regarding sensitive content (Question Four) and emphasises why exemptions do not completely eliminate the risk of having sensitive content in a collection. The proposed workflows presented in Chapter 7 WORKFLOWS, if implemented, would ensure that these considerations are consistently undertaken.

The next topic of discussion involves the options and choices available regarding resources, tools, and workflows.

As has been previously mentioned, the digital forensic methods will add processing time. There are some options that are available to improve this situation. The first option is to acquire better hardware that is faster, has more processing power, and can perform more tasks. This is the best option, but of course, it comes at a cost. If this is not a viable option due to resource limitations, then when and how data are processed must be considered more carefully. Without adequate hardware, processing large amounts of data in batches may not be viable. Instead, it may be more efficient to process data as they are acquired if there are enough resources to allocate to the task. This will reduce the size of each processing task which will lessen the time taken and minimise stress on the system as fewer modules may be operating, depending on the type of software used.

Note that a significant increase to processing time will be determined by the analysis of output which must be performed by staff. Training, education, and staff allocation are factors that will determine the impact to overall processing time. The ability to identify anomalies, false-positives, and key information are important skills that are developed over time with the aid of training. It is worth noting that the knowledge from departments that are making legal and ethical decisions will also impact the time it takes for processing to proceed past certain steps.

If the “Other” material poses no risk to any “Target” data, then it can be stored for processing at a more convenient time. The level of processing is determined by the type of data. If the data are purely text-based, then a digital forensic tool that focuses primarily on text should be used as a complete forensic package is not necessary. When the data are made up of multiple data types, but the point of interest is primarily text-based, the initial analysis may provide incentive to further process the remaining data. It may then be decided to utilise a digital forensic package that has the capability to process image, video, peripheral data, communications, and network data, for example. Each case is unique, and decisions must be made on the level of processing required whilst factoring time and resource constraints.

An important point to reiterate is that the discovery and handling of sensitive data is not only for risk mitigation, but it can also be beneficial for a collection. These data can provide information that gives extra context to collection items and help in identifying system environment information to aid in the development of emulated environments as discovered in the experiments presented in Chapter 6 DIGITAL FORENSICS – SENSITIVE DATA. The added benefit of sensitive data discovery and handling will improve various areas of the digital preservation process. By knowing what types of sensitive data exist and how they can impact a collection allows donor investigations to be conducted more thoroughly. Anything that is not covered in the donor process can be mitigated with proper checks and procedures to not allow data to be processed without having been properly investigated as shown in Chapter 7 WORKFLOWS. This relates to research Question one regarding how improvements can be made in data gathering capabilities, offering benefits other than simply the discovery and redaction of sensitive data, as discussed in Section 6.2 Collection Institution Relevance regarding the relevance of digital forensic tools in collection institutions.

Addressing research Question two, each solution proposed significantly increases data gathering capabilities and allows data to be discovered in obscure locations that would not be possible without such digital forensic tools. There are many options that have been identified through looking into workflows of other institutions and the tools that were tested for this study have provided insight into their potential. There are often alternatives that perform the same tasks, giving collection institutions a choice in what they use. There are free, open-source solutions, and there are propriety commercial products. The tools tested in this study, Bulk_extractor and Autopsy (TSK), are open-source, with online documentation and support. Autopsy is a forensic package with many interesting features useful to both digital preservation and digital forensic analysts. EnCase and FTK are alternatives to this, and they may provide a feature that is desired by certain institutions. For example, EnCase promote their mobile investigator feature for use on mobile devices; however, EnCase is a proprietary solution. Solutions such as BitCurator exist where many of the tools and techniques mentioned are packaged together and modified for preservation purposes.

There are more solutions and there will be new developments over time. There will always be choices that can be made to suit any situation. No one tool is being recommended over another and the purpose of experimenting with the tools in this study was to provide an overview of their potential with a focus on the output and results. Some considerations that should be made when selecting a tool consist of:

- Resource limitations
- Online documentation
- Online support
- Features
- Ease of use
- Credibility (trust)

This also ties into staff training. The results from the questionnaire revealed that there was a case where the adoption of new tools and methods was unwanted due to staff training requirements. This has been a consideration when designing workflow enhancements and suggestions, but it is difficult to provide solutions that do not require some form of training.

Depending on staff availability, there may be viability in providing training to one staff member who can then train others on the job. Operating the software and analysing the output are two different tasks requiring a different skillset. Training staff to use tools can be achieved without the use of many resources, but there should be dedicated training on how to analyse the output. Staff can then learn from one another and is something that can be achieved in processing down-time.

The majority of the institutions surveyed, through their public information and responses to the questionnaire, are not performing digital forensics on a level which enables the ability to discover and handle sensitive data. This has been determined based on whether institutions have acquired the necessary equipment, how they use said equipment, and their overall preservation process which indicates whether sensitive data discovery is being performed. If the means to discover sensitive data are not available, one cannot safely assume the task is being performed correctly. Basic metadata extraction may be utilised using basic features of the operating system and other tools; however, these methods are not effective enough.

There will be a learning curve, but the benefits of adopting digital forensic tools and methods, providing the potential to remove and mitigate future issues, should be considered against the cost of training and the purchasing of equipment and software. Ensuring the data published in collections is authentic and accurate is invaluable when compared to how a single mistake can jeopardise the reputation of an institution and potentially its associated partners. The implementation of the suggested enhancements can be done at the pace of the institutions choosing. Every institution is at a different level of maturity. Some institutions are still in their infancy regarding their preservation needs. They may not have a large intake of digital material in need of preservation, but in time, this will increase.

Therefore, the workflows provided are flexible. The choice in design and notation is made by an institution to construct a workflow that best suits its requirements. The ultimate goal is to provide an accurate visual representation of the preservation workflow to achieve full transparency. Transparency allows users and donors to see exactly how their data will be handled, as well as allowing peer institutions to learn from one another, eventually leading to the widespread betterment of digital preservation.

# 9 CONCLUSION and RECOMMENDATIONS

The initial investigations of this thesis focused on the individual fields of digital preservation and digital forensics. A more personal and direct approach was taken for digital preservation to gain an understanding of what is lacking within a typical preservation workflow for Australian collection institutions. As for digital forensics, literature and online resources were researched to identify the many different methods and tools unique to the field.

This research faced limitations in the data gathering aspects and the sample size of the participating Australian institutions. The targeted sample size was every state and national library of Australia, these were selected as they represent their state, territory, or nation. The objective was to solve the problems at leading institutions so they can set a better example for smaller institutions. Not all institutions accepted participation, and some were unable to participate as their digital preservation maturity level was not developed enough to provide sufficient information. Some institutions had to withdraw from participation and the archives that did participate provided insufficient information and where therefore omitted.

Other limitations throughout this thesis were imposed intentionally. The data from the U.S institutions were restricted to easily accessible, publicly available, transparent sources of data. This required full access without any payment or membership requirements and a detailed workflow displaying the steps and processes of the workflow and the supporting tools. The experimentation of software was done on freely available tools that either had a strong presence among the institutions investigated or were competitors to popular proprietary alternatives to sympathise with budgetary and staff resource restrictions.

The final limitation imposed was on the enhanced workflows. Care was taken to ensure the enhancements can be amended to existing workflows, therefore, no changes were to be made to existing procedures. Improvements may be made at the discretion of the institution. The main contributions include a more informative investigation into the donor to better anticipate the presence of sensitive data, as well as the ability to discover and then handle sensitive data through to storage and access stages of the workflow. The workflow design and notation were kept simple and flexible, modularised to be better suited for adoption into existing workflow models.

This thesis highlights the similarities and differences between applying digital forensics to a criminal matter and digital preservation. In a criminal case, digital forensics is used to discover data which is then correlated into evidence in order to prosecute and uncover misconduct. The very nature of the criminal aspect separates it from digital preservation as the data processed in a forensic investigation serves no purpose once a case is closed. Digital preservation is concerned with long-term goals and providing meaningful access to their data. The caution required when handling data is a shared concern for both fields. Caution is required when handling data in an investigation if the data is to be accepted as evidence, much like a collection institution must handle its data with care to ensure the accuracy, completeness, and integrity of the data is not impacted by irreversible change.

With this combined research, it revealed where digital preservation processes could be improved and how digital forensics could accommodate this, answering the first research question proposed. Whilst digital preservation has much to gain from digital forensics, digital preservation methodologies may be viable for adoption by the digital forensics community. This includes how data are treated and cared for long-term, which may be viable should an investigation be re-opened. With the process of keeping change history and other related metadata on who and what has handled the data during its lifespan, should a new investigator be required, they have the data needed to better handle issues that may arise. In this case, if considerable time has passed, the media in which the data are stored may have degraded if not taken proper care of as a digital preservation collection would ensure.

The collaborative nature of digital preservation could influence the design and documentation of digital forensic tools and methods to consider and acknowledge other fields where these tools and methods may be beneficial.

With this initial investigation, it was clear that sensitive data was forming the main focal point of the study. Digital preservation has two foci. One is the digitisation and preservation of content from physical media, legacy devices, and historical recordings such as paintings and other hand-written material. The other focus, which is becoming more prominent, is born-digital data. Hardware is always advancing as is how it is used, meaning there will be a larger emphasis on born-digital data. For example, mechanical hard drives are being replaced by solid state drives which are rapidly becoming more affordable with larger storage capacity. Solid state drives are also advancing with the M.2 form factor, changing their physical design and how they are connected. How collection institutions approach these new form factors in a preservation environment may require new tools and setups. Traditional devices that allow hard disk drives and 2.5-inch solid state drives to be connected via USB will not be compatible with PCIe and M.2 form factors without an adaptor.

As such, this study is concerned with born-digital data and future technological issues. Therefore, the access and use of legacy hardware and media is not a concern in this thesis. Current media and technology are slowly but surely replacing legacy media within collections. This means the needs and requirements for preserving modern data will grow. Hopefully collection institutions can process their backlogs of legacy media (mountains of floppy disks sitting in storage) before modern media starts to pile up.

From this the emphasis on sensitive data emerges. Our digital footprint is bigger than ever and as technology becomes more approachable and accessible, more people are going to use technology for various novel activities and purposes. This means more novice users that are not aware of the digital footprint they are creating. With this, the amount of data created is going to increase significantly and exponentially. This was shown in the experiments conducted which resulted in two disk images, one based on a single user directory and the other an entire hard drive. Both images were of the same size and produced similar amounts of data. One directory was able to produce the same amount as an entire hard drive, not just because of the size similarity, but due to how much use the personal computer had in which the directory came from.

This raises the following questions:

- What types of sensitive data are within the collection?
- How can these data be accessed?
- What are the implications of finding or not finding sensitive data?
- What is to be done when sensitive data are found?
- What are the legal concerns?
- What are the ethical concerns?

Some institutions may see this as an unprecedented event and not see a threat; however, this does not rule out the possibility that these threats already exist within a collection. Without the proper digital forensic techniques, it is impossible to accurately determine this.

With the establishment of sensitive data being a critical issue, the focus was then on how to discover if adequate methods were being performed to discover and handle these data. Through the investigation of public material and direct communication, it was revealed which institutions were performing such tasks, and which were not. From this it was discovered that workflows, being the key source of showing how an institution performs, were lacking in most areas.

There are two possible reasons for why the workflows were lacking. One is because the digital preservation workflow did not include certain tasks, meaning the collection institution was not performing them. Secondly, the institution may potentially not be visualising everything that is being performed, which may include intentional omission. Institution maturity is a likely factor in this, tied to the requirements and demands of digital preservation per institution.

It was clear at this point that collection institutions performing digital preservation on a large scale would benefit from adopting digital forensic tools and methods. This would enable sensitive data discovery capabilities, and in turn, improve the handling of sensitive data in institutions that are not already doing so. It also revealed that workflows in general needed to be improved. There were some workflows from international institutions which were designed well and met most of the criteria specified; however, even the exemplary workflows had room for improvement. These criteria were:

- Donor agreement
- Sensitive data discovery
- Sensitive data handling.

Meeting all these criteria is necessary to fully achieve a transparent workflow from which other institutions can learn and instil trust to donors. Many collection institutions are parts of a collaborative group, such as the National and State Libraries Australia (NLSA), a body for Australia’s national, state and territory libraries (NSLA, 2020). Therefore, the idea of full transparency is important in working towards standardisation.

From this the main contributions and discoveries of this study are brought forth.

As digital forensics is the suggested solution to enhance digital preservation processes, an investigation into how this can be achieved was necessary, specifically the tools and methods used by other institutions. After gathering data on the tools being used from a set of institutions, it led to the acquisition and experimentation of two specific tools which cover a wide variety of functionality as there are many alternatives that perform the same tasks. This investigation and eventual experimentation of the selected tools was essential in answering the second research question on how digital forensic tools and techniques can be implemented to resolve data gathering issues.

The results revealed much about the potential and benefits these tools have to offer. The experiments were conducted with a digital preservation perspective which differs from the perspective required by a forensic analyst when investigating a criminal case. In this regard, how sensitive data can give additional context to collection items was explored. For example, if a collection were based on an iconic figure and the institution had their personal files in possession, there could be hidden data that reveal interesting information about them. Anything from personal interests, hobbies, or views on life, could all be derived from their digital footprint. A more pessimistic view is that this information could reveal degrading information that tarnishes the reputation of that iconic figure.

The results revealed that having the capability of discovering data that could not be obtained by other methods, severely increases the chance of finding data of a sensitive nature, accompanied by risk. This ushers in issues surrounding legality and ethics, but these can be mitigated and avoided with appropriate action such as in-depth donor agreements and accession record keeping. Without this discovery, collection institutions may have data lying dormant in their repositories which still have a purpose to serve. They may also be providing access to information that is not complete and therefore not accurate. The extent of sensitive data discovery is pertinent to the remaining two research questions which address the consideration of legal implications, despite collection exemptions, and ethical decision making in the same regard.

The accuracy of information and the message it delivers is something that can be completely altered with the slightest change. This has been seen in doctored images and hoaxes throughout the Internet where a simple edit can completely change the nature of an image (Hart and de Vries, 2017). The same can be said for the data residing in collection institutions. If data were derived from a device and were easy to discover in a convenient location, any further processing might not be seen as necessary. What happens to the media and the remaining data is typically not addressed in public information and workflows. It is unlikely the media is processed any further once the desired data are extracted. This is supported from seeing institutional workflows that do not reveal sensitive data discovery and handling. There may be data in the discards that provide a more complete picture of what is being preserved and it may improve it, change it, or make it unworthy for access.

The laws and implications sensitive data bring to Australian institutions have been investigated. This also includes ethics as these areas are often met with exemption from the surrounding law, specifically an institution’s collection. In the event of an exemption, this is something that should not be ignored. This may be classified as a grey area due to there being cases of defamation where exemptions have been overruled. Loopholes exist in laws, and these can be exploited, even in collection institutions. Therefore, it is important to investigate the relevant laws and discuss how they should be treated. The Australian laws that have been investigated have been deemed suitable guidelines, even if they are neither obligatory nor enforceable in most cases. Using these laws as guidelines can protect an institution should their exemption fail and will further allow them to maintain an ethical stature, which will mitigate negative views from the public and of course, donors.

As there are limitations each institution must face, it is not feasible to try to solve all the problems at once. Instead, the solution is to improve and enhance workflows so that it may cause a cascading effect where the other institutions start to learn from each other as their preservation needs increase. It is recommended to add digital forensic methods to workflows where they do not exist and improve supporting processes such as donor agreements and how sensitive data are handled once discovered. Any solutions provided are mindful of existing procedures and aim to cause as little disruption as possible. There are of course solutions that involve cost and training, but there are other options such as open-source solutions. There will still be training requirements which may incur a cost and additional processing time will be required to analyse the output of processed sensitive data.

The benefits of digital forensics to digital preservation institutions have been presented, so too have the risks should these methods not be used. The tools and methods shared across multiple institutions, and those that are used exclusively, have been recorded as data and presented in a way that shows the overall influence certain tools have. This method led to other discoveries such as why certain tools were used over others and why certain institutions were still using tools that have been superseded; ultimately leading to the choice in which tools were tested.

The potential of these tools has been explored with many small-scale examples and results. Many examples, both real and hypothetical, have been provided with various levels of severity. The severity may be relative, as no two institutions are identical, but this is not a static variable. Therefore, it is unwise to ignore the threats just because an institution is small and does not have a reason for having a dedicated preservation or digital forensics workflow. Some institutions may remain this way for quite some time, but when they are ready, it is best to approach the changes correctly and with good example to follow, thus being the ultimate goal of this study.

The culmination of the research, experiments, and eye-opening experiences has all led to the forensically enhanced digital preservation workflows. These workflows have been carefully constructed to meet all criteria and consideration discussed throughout this study. The workflows provide a framework to be implemented and configured based on the flexible needs of collection institutions. These workflows address the initial, processing, and final stages of the preservation workflow, in accordance with all research questions, addressing:

- Where improvements can be made
- The increase in data gathering capabilities
- How digital forensic tools achieve this
- Donor agreement and interview improvements to better prepare for sensitive data
- Decision making and checks to remove or mitigate legal and ethical risk.

The enhancements will allow collection institutions to be confident in the data they publish being authentic, complete, and accurate. The reduction and removal of risk, and the added benefits of thorough data gathering capabilities will be essential in progressing collection institutions towards the future. Sensitive data discovery, the decision-making to handle such data, and the procedures to follow at the beginning of ingest are all improvements that will prepare these institutions for larger intakes of born-digital data.

With this study, progress can be made by evaluating where these institutions can improve, not only with the tools they use or the workflows they follow, but also by adding to their processes, because it is apparent that there are some important processes missing. These range from dedicated preservation processes to the full extent of digital forensic investigation. The need for such additions may not be necessary at the current maturity level of certain institutions, but digital preservation is ever-changing and will continue to increase in demand. It is best to prepare for future issues and prevent them, rather than wait and deal with the issues that arise.

Some collection institutions in Australia may be in their infancy regarding digital preservation, but there are clearly some institutions that are leading and the example they set may be followed by the institutions that are progressing slower. Progression is not based on competence, but by the demand on the institutions, the resources available, and the level of importance placed on preserving our digital history.

# 10 FUTURE WORK

There is work to be done that can aid in achieving the goals of this research as well as to help institutions adopt these enhancements. Awareness, being a key factor, is not only important within the preservation community, but it also something the digital forensics community should consider. There is a gap between the two disciplines and any future work should aim towards bridging this gap. The following paragraph highlights the gap, the solution, and the issue:

> *"The intersection between digital forensics and archives can be characterized as a 'trading zone' that resides between different streams of activity. Individuals and groups can agree to use a common set of terms, concepts and methods in order to share ideas and coordinate their work, even if they still hold dramatically different worldviews, values or assumptions of their own responsibilities. It is likely that fundamental elements of digital forensics language and practice will ultimately become so embedded in the archival enterprise that archivists no longer perceive them as being borrowed from elsewhere; they will simply be part of what archivists do. As archivists develop new methods and tools that are based on forensics building blocks, hopefully they will also make contributions to the field of digital forensics that it can ultimately adopt as established practice. However, it is also likely that the frontiers of digital forensics and archival research will continue to develop independently, based on distinct values, mandates and constraints. There is the potential for creative and well-informed translation work across the two streams for many years ahead."* (Lee, 2012)

If digital forensics tools were more approachable for non-forensic disciplines, the preservation community would find them easier to adopt and make use of the tools and methods that are available. Nonetheless, more effort may be made within the preservation community to explore these areas of digital forensics and be willing to learn and embrace them. This study has shown the potential of these tools and methods that are well suited and beneficial towards collection institution goals. These benefits are achieved by significantly improving data gathering capabilities for both sensitive and non-sensitive data, improving collection accuracy, authenticity, and completeness, as well as allowing meaningful access. The exposure of risk that may already exist and that will be more prominent with the use of the digital forensic tools and methods should promote action towards the mitigation or prevention of future legal and ethical issues in the ever-changing political landscape of privacy. Therefore, policies need to be re-evaluated and amended to accommodate this and to achieve a higher level of transparency. Based on the findings of this research, there are several options the digital preservation community could consider. Workflow tools can be re-purposed to better suit the flexible and unique nature of digital preservation. Modular design can be put into practice by having the means to digitally construct, modify, and share workflows.

Collaborative groups and communities can work towards providing a better means for peer institutions to standardise their protocols, procedures, and methods. Standardisation can help in the long term when the smaller institutions are catching up. These institutions can learn directly from their peers, but by having a central, standardised, and trusted parent institution, changes and improvements on larger scale can be achieved. It is much easier to make changes based on an existing model, rather than creating something new. Therefore, this will allow institutions to modify the standard approach to meet their needs.

The digital preservation community may take it upon themselves to address the gap by taking the online resources and documentation accompanying digital forensic tools and modifying them to their perspective. This is after all one of the issues regarding the tools that are available. Although these tools can perform certain tasks, for example, task **X** and task **Y**, the manuals and online documentation may only describe how these tasks are performed from a criminology point of view (**X**). Whereas **Y** can be tailored and utilised for preservation goals, but without the documentation to explain this, these solutions may not be discoverable. Without someone to test this software and publish the results from a digital preservation collection's perspective, how are others meant to discover the capabilities of such tools?

All future work should be done with the goal to bring these two disciplines closer together so they can strengthen one another in a more user friendly and collaborative approach. Whilst initial efforts should be made to improve collection institutions from within, collaboration and countrywide improvement should be the next step. International collaboration between digital preservation communities will be harder to achieve with many different laws and ethical views across jurisdictions, but this does not mean collection institutions cannot learn from one another. The future is digital, larger, and riskier; now is the time to prepare.

# 11 REFERENCES

AccessData, 2018. Forensic Toolkit [WWW Document]. URL https://accessdata.com/products-services/forensic-toolkit-ftk (accessed 11.21.18).

AccessData, 2017. FTK Imager [WWW Document]. AccessData. URL https://accessdata.com/product-download/ftk-imager-version-4.2.0 (accessed 12.3.18).

Aguilar-Saven, R.S., 2004. Business process modelling: Review and framework. International Journal of production economics 90, 129–149.

ArchivesSpace, 2020. History | ArchivesSpace. URL https://archivesspace.org/about/history (accessed 4.22.20).

Artefactual, 2019. Archivematica: open-source digital preservation system [WWW Document]. URL https://www.archivematica.org/en/ (accessed 6.17.19).

Ashley, L.J., Misic, M., 2019. Digital Preservation Capability Maturity Model (DPCMM): Genesis and Practical Uses, in: Diverse Applications and Transferability of Maturity Models. IGI Global, pp. 152–167.

ATSILIRN, 2012. ATSILIRN - Aboriginal and Torres Strait Islander Library and Information Resource Network [WWW Document]. URL http://atsilirn.aiatsis.gov.au/index.php (accessed 6.6.18).

Australian Government, 2022. Freedom of Information Act 1982 [WWW Document]. URL https://www.legislation.gov.au/Details/C2022C00036/Html/Text, http://www.legislation.gov.au/Details/C2022C00036 (accessed 2.1.22).

Australian Government, 2018. Privacy Act 1988 [WWW Document]. URL https://www.legislation.gov.au/Details/C2018C00292 (accessed 3.13.18).

Awad, A., Kadry, S., Lee, B., Maddodi, G., O'Meara, E., 2016. Integrity Assurance in the Cloud by Combined PBA and Provenance, in: 2016 10th International Conference on Next Generation Mobile Applications, Security and Technologies (NGMAST). Presented at the 2016 10th International Conference on Next Generation Mobile Applications, Security and Technologies (NGMAST), pp. 127–132. https://doi.org/10.1109/NGMAST.2016.15

Babich, N., 2018. The 12 Do's and Don'ts of Web Design. Adobe Blog. URL https://theblog.adobe.com/12-dos-donts-web-design-2/ (accessed 6.18.19).

Barrett, C., 2017. Digital forensics tools and methodologies in archival repositories.

Bartliff, Z., Kim, Y., Hopfgartner, F., Baxter, G., 2020. Leveraging digital forensics and data exploration to understand the creative work of a filmmaker: A case study of Stephen Dwoskin’s digital archive. Information Processing & Management 57, 102339.

Basis Technology, 2018a. The Sleuth Kit (TSK) & Autopsy: Open Source Digital Forensics Tools [WWW Document]. URL https://www.sleuthkit.org/ (accessed 7.11.18).

Basis Technology, 2018. Autopsy [WWW Document]. URL https://www.sleuthkit.org/autopsy/ (accessed 10.3.18).

Basis Technology, 2018b. Autopsy User Documentation: Content Viewer [WWW Document]. URL https://sleuthkit.org/autopsy/docs/user-docs/4.8.0/content_viewer_page.html (accessed 11.19.18).

Basis Technology, 2018c. Autopsy User Documentation: Autopsy User’s Guide [WWW Document]. URL https://sleuthkit.org/autopsy/docs/user-docs/4.8.0/index.html (accessed 11.19.18).

Basis Technology, 2018d. Autopsy User Documentation: Communications Visualization Tool [WWW Document]. URL https://sleuthkit.org/autopsy/docs/user-docs/4.8.0/communications_page.html (accessed 11.19.18).

Basis Technology, 2018e. Autopsy User Documentation: Timeline [WWW Document]. URL https://sleuthkit.org/autopsy/docs/user-docs/4.8.0/timeline_page.html (accessed 11.20.18).

Baucom, E., 2019. A brief history of digital preservation. Digital preservation in libraries: Preparing for a sustainable future 3–19.

Bishop, B., 2018. James Gunn fired from Guardians of the Galaxy 3 over offensive tweets - The Verge [WWW Document]. The Verge. URL https://www.theverge.com/2018/7/20/17596452/guardians-of-the-galaxy-marvel-james-gunn-fired-pedophile-tweets-mike-cernovich (accessed 3.5.19).

BitCuractor Consortium, 2018. Workflows [WWW Document]. URL https://bitcuratorconsortium.org/workflows (accessed 7.4.18).

BitCurator, 2018. BitCurator [WWW Document]. URL http://bitcurator.net/ (accessed 8.15.16).

Bonta, R., 2018. California Consumer Privacy Act (CCPA) [WWW Document]. Rob Bonta Attorney General. URL https://oag.ca.gov/privacy/ccpa (accessed 2.1.22).

Business Process Modelling [WWW Document], 2018. . Wikipedia. URL https://en.wikipedia.org/wiki/Business_process_modeling (accessed 12.3.18).

CCSDS, 2012. Reference Model for an Open Archival Information System (OAIS) 135.

Chassanoff, A., Woods, K., Lee, C.A., 2016. Digital Preservation Metadata Practice for Disk Image Access, in: Digital Preservation Metadata for Practitioners. Springer, pp. 99–109.

Cho, J., Chen, I., 2018. PROVEST: Provenance-Based Trust Model for Delay Tolerant Networks. IEEE Transactions on Dependable and Secure Computing 15, 151–165. https://doi.org/10.1109/TDSC.2016.2530705

Cochrane, E., Tilbury, J., Stobbe, O., 2018. Adding emulation functionality to existing digital preservation infrastructure. Journal of Digital Media Management 6, 255–264.

COPTR, 2021. Community Owned Digital Preservation Tool Registry (COPTR) [WWW Document]. URL https://coptr.digipres.org/Main_Page (accessed 12.16.20).

Crazy Egg, 2018. Effective Web Design | 8 Do’s and Don’ts [WWW Document]. The Daily Egg. URL https://www.crazyegg.com/blog/the-dos-and-donts-of-effective-web-design/ (accessed 6.18.19).

Cunningham, A., 2008. Digital curation/digital archiving: A view from the National Archives of Australia. The American Archivist 71, 530–543.

DCMI, 2018. Dublin Core Metadata Initiative [WWW Document]. URL http://dublincore.org/ (accessed 9.3.18).

Dekker, A., 2018. Collecting and conserving net art: moving beyond conventional methods. Routledge.

DeRidder, J.L., Helms, A.M., 2016. Intake of digital content: Survey results from the field. D-Lib Magazine 22.

Dietrich, D., Adelstein, F., 2015. Archival science, digital forensics, and new media art. Digital Investigation 14, S137–S145. https://doi.org/10.1016/j.diin.2015.05.004

Dietrich, D., Kim, J., McKeehan, M., Rhonemus, A., 2016. How to Party Like it’s 1999: Emulation for Everyone. The Code4Lib Journal.

Dobreva-McPherson, M., Kim, Y., Ross, S., 2013. Automated Metadata Generation. DCC | Digital Curation Reference Manual ISSN 1747-1524 http://www.dcc.ac.uk/resources/curation-reference-manual/completed-chapters/automated-metadata-extraction.

Doctor, D., 2007. The new uniform defamation laws [WWW Document]. Arts Law. URL https://www.artslaw.com.au/articles/entry/the-new-uniform-defamation-laws/ (accessed 3.20.18).

Dollar, C., Ashley, L., 2015. Digital Preservation Capability Maturity Model (DPCMM).

Duranti, L., 2016. Archival strategies for a constantly connected society. Seminário Serviços de Informação em Museus 193–201.

Duranti, L., Rogers, C., 2016. Trust in records and data online. Integrity in Government through Records Management. Essays in Honour of Anne Thurston 203–214.

Eastlake 3rd, D., Jones, P., 2001. US secure hash algorithm 1 (SHA1).

Eckard, M., Hagen, A., 2018. 401.2 Revamping the “Difficult (Potentially)” but “Mostly Good” and “Pretty Smooth” Removable Media Workflow at the Bentley Historical Library. p. 5.

Educopia, 2018. OSSArcFlow | Educopia Institute [WWW Document]. URL https://educopia.org/OSSArcFlow/ (accessed 2.19.19).

Faisal, C.N., Gonzalez-Rodriguez, M., Fernandez-Lanvin, D., de Andres-Suarez, J., 2016. Web design attributes in building user trust, satisfaction, and loyalty for a high uncertainty avoidance culture. IEEE Transactions on Human-Machine Systems 47, 847–859.

Ferme, V., Lenhard, J., Harrer, S., Geiger, M., Pautasso, C., 2017. Workflow management systems benchmarking: unfulfilled expectations and lessons learned, in: Proceedings of the 39th International Conference on Software Engineering Companion. IEEE Press, pp. 379–381.

Fileinfo, 2017. PLIST File Extension [WWW Document]. URL https://fileinfo.com/extension/plist (accessed 11.19.18).

ForensicsWiki, 2017. Raw Image Format [WWW Document]. URL https://www.forensicswiki.org/wiki/Raw_Image_Format (accessed 12.3.18).

Freed, N., Kucherawy, M., 2019. MIME [WWW Document]. iana. URL https://www.iana.org/assignments/media-types/media-types.xhtml

Gallinger, M., Bailey, J., Cariani, K., Owens, T., Altman, M., 2017. Trends in Digital Preservation Capacity and Practice: Results from the 2nd Bi-Annual National Digital Stewardship Alliance Storage Survey. D-Lib Magazine 23.

Garfinkel, S., 2013. Digital media triage with bulk data analysis and bulk_extractor. Computers & Security 32, 56–72. https://doi.org/10.1016/j.cose.2012.09.011

Gartner, R., Lavoie, B., 2013. Preservation Metadata (2nd Edition). Digital Preservation Coalition.

Gengenbach, M., 2012. The way we do it here": Mapping digital forensics workflows in collecting institutions.". Unpublished master's thesis, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina.

GIDA, 2019. CARE Principles of Indigenous Data Governance [WWW Document]. Global Indigenous Data Alliance. URL https://www.gida-global.org/care (accessed 7.20.20).

Gidney, K., 2016. Restricted! URL https://www.nla.gov.au/blogs/behind-the-scenes/2016/08/24/restricted (accessed 3.19.18).

Glăveanu, V.P., 2014. Distributed creativity: Thinking outside the box of the creative individual. Springer.

Global Negotiator, n.d. What is Commercial in confidence? Definition and meaning. Dictionary of International Trade. URL https://www.globalnegotiator.com/international-trade/dictionary/commercial-confidence/ (accessed 3.19.18).

Greene, M., Meissner, D., 2005. More product, less process: Revamping traditional archival processing. The American Archivist 68, 208–263.

Grigorova, K., Mironov, K., 2018. Conversion of business process models using workflow patterns, in: 2018 5th International Conference on Control, Decision and Information Technologies (CoDIT). IEEE, pp. 763–766.

Guerra, P.C., Nalon, R., Assunção, R., Meira Jr, W., 2017. Antagonism also flows through retweets: The impact of out-of-context quotes in opinion polarization analysis, in: Eleventh International AAAI Conference on Web and Social Media.

Guidance Software, 2018. EnCase Forensic Software [WWW Document]. URL https://www.guidancesoftware.com/encase-forensic (accessed 11.21.18).

Han, M.-J.K., 2016. Establishing sustainable and scalable workflows for cataloging and metadata services. Library Management 37, 308–316. https://doi.org/10.1108/LM-04-2016-0031

Hart, T.R., 2015. Metadata Standard for Future Digital Preservation. Flinders University-Adelaide Australia.

Hart, T.R., de Vries, D., 2017. Metadata Provenance and Vulnerability. Information Technology and Libraries 36, 24–33.

Hart, T.R., de Vries, D., Mooney, C., 2019. Australian Law Implications on Digital Preservation. Presented at the iPres 2019, Amsterdam. https://doi.org/10.17605/OSF.IO/EZ6FQ

Hartig, O., Zhao, J., 2010. Publishing and consuming provenance metadata on the web of linked data, in: International Provenance and Annotation Workshop. Springer, pp. 78–90.

Harvard Library, 2018. File Information Tool Set (FITS) [WWW Document]. URL http://projects.iq.harvard.edu/fits (accessed 8.15.16).

Harvard University, 2011. Differential Privacy [WWW Document]. URL https://privacytools.seas.harvard.edu/differential-privacy (accessed 2.21.22).

Harvey, R., 2015. The last decade of digital preservation: a personal view from Australia. Preservation, Digital Technology & Culture 44, 22.

Harvey, R., Mahard, M., 2013. Mapping the preservation landscape for the twenty-first century. Preservation, Digital Technology & Culture 42, 5–16.

Hasan, R., Sion, R., Winslett, M., 2009. The Case of the Fake Picasso: Preventing History Forgery with Secure Provenance., in: FAST. pp. 1–14.

Hofman, D., Duranti, L., How, E., 2017. Trust in the Balance: Data Protection Laws as Tools for Privacy and Security in the Cloud. Algorithms 10, 47.

Huan, L., 2006. Uniform Defamation Laws 2006. Stephens Lawyers & Consultants. URL http://www.stephens.com.au/the-uniform-defamation-laws-2006/ (accessed 3.22.18).

Ingram, D., Henshall, P., 2019a. Chapter 69: Defamation - what you cannot do [WWW Document]. URL http://www.thenewsmanual.net/Manuals%20Volume%203/volume3_69.htm (accessed 3.20.18).

Ingram, D., Henshall, P., 2019b. Defamation in Australia [WWW Document]. URL http://www.thenewsmanual.net/Resources/medialaw_in_australia_02.html (accessed 3.20.18).

Jaillant, L., 2022. How can we make born-digital and digitised archives more accessible? Identifying obstacles and solutions. Archival Science. https://doi.org/10.1007/s10502-022-09390-7

John, J.L., 2012. Digital Forensics and Preservation. DPS Technology Watch Report. http://dx.doi.org/10.7207/twr12-03

Jones, M., Beagrie, N., 2001. Preservation management of digital materials: a handbook. British Library London.

Khanna, N., Mikkilineni, A.K., Delp, E.J., 2009. Research and Technology - Forensic Camera Classification: Verification of Sensor Pattern Noise Approach - January 2009 [WWW Document]. FBI. URL https://www.fbi.gov/about-us/lab/forensic-science-communications/fsc/jan2009/research/2009_01_research01.htm (accessed 8.10.16).

Kim, Y., Ross, S., 2012. Digital forensics formats: seeking a digital preservation storage container format for web archiving. International Journal of Digital Curation 7, 21–39.

Kingston, 2020. Understanding SSD Technology: NVMe, SATA, M.2 [WWW Document]. Kingston Technology Company. URL https://www.kingston.com/en/community/articledetail/articleid/48543 (accessed 10.6.20).

Kirschenbaum, M.G., Ovenden, R., Redwine, G., Donahue, R., 2010. Digital forensics and born-digital content in cultural heritage collections, CLIR publication. Council on Library and Information Resources, Washington, D.C.

Knight, G., 2012. The Forensic Curator: Digital Forensics as a Solution to Addressing the Curatorial Challenges Posed by Personal Digital Archives. International Journal of Digital Curation 7. https://doi.org/10.2218/ijdc.v7i2.228

Kohn, M.D., Eloff, M.M., Eloff, J.H.P., 2013. Integrated digital forensic process model. Computers & Security 38, 103–115. https://doi.org/10.1016/j.cose.2013.05.001

Küng, S., 2018. grepWin: Regular expression search and replace for Windows [WWW Document]. Stefans Tools. URL https://tools.stefankueng.com/grepWin.html (accessed 11.1.18).

Kussmann, C., Alliance, D.S., Graham, W., Atkins, W., Reich, A., 2020. 2019 Levels of Digital Preservation Matrix. OSF.

Lampert, C., Vaughan, J., 2018. Preparing to Preserve: Three Essential Steps to Building Experience with Long-Term Digital Preservation 13.

Langley, S., 2020. Planning for the End from the Start: An Argument for Digital Stewardship, Long-Term Thinking and Alternative Capture Approaches for Digital Content, in: Digital Cultural Heritage. Springer, pp. 209–237.

Langley, S., 2018. Digital Preservation Should Be More Holistic: A Digital Stewardship Approach. American Library Association.

Larson, E., 2020. Big Questions: Digital Preservation of Big Data in Government. The American Archivist 83, 5–20.

Lazorchak, B., 2015. Digital Forensics and Digital Preservation: An Interview with Kam Woods of BitCurator. | The Signal: Digital Preservation [WWW Document]. URL https://blogs.loc.gov/digitalpreservation/2015/05/digital-forensics-and-digital-preservation-an-interview-with-kam-woods-of-bitcurator-2/ (accessed 8.10.16).

LeClere, E., 2019. Breaking rules for good? How archivists manage privacy in large-scale digitisation projects. Archives and Manuscripts 0, 1–20. https://doi.org/10.1080/01576895.2018.1547653

Lee, C., 2012. Archival application of digital forensics methods for authenticity, description and access provision. Comma 2012, 133–140. https://doi.org/10.3828/comma.2012.2.14

Lee, C.A., 2018. Computer-Assisted Appraisal and Selection of Archival Materials, in: 2018 IEEE International Conference on Big Data (Big Data). IEEE, pp. 2721–2724.

Lee, H.C., Palmbach, T., Miller, M.T., 2001. Henry Lee's crime scene handbook. Academic Press.

Library of Congress, n.d. Metadata Encoding and Transmission Standard (METS) Official Web Site | Library of Congress [WWW Document]. URL http://www.loc.gov/standards/mets/ (accessed 8.30.18a).

Library of Congress, n.d. PREMIS: Preservation Metadata Maintenance Activity (Library of Congress) [WWW Document]. URL https://www.loc.gov/standards/premis/ (accessed 8.30.18b).

Loechner, J., 2016. 90% Of Today's Data Created In Two Years 12/22/2016 [WWW Document]. MediaPost. URL https://www.mediapost.com/publications/article/291358/90-of-todays-data-created-in-two-years.html (accessed 5.20.19).

Loukides, G., Standen, L., Fai, M., 2020. Why Australian businesses should care about the California Consumer Privacy Act [WWW Document]. Gilbert + Tobin. URL https://www.gtlaw.com.au/knowledge/why-australian-businesses-should-care-about-california-consumer-privacy-act (accessed 2.1.22).

Lucidchart, 2019. What is a Swimlane Diagram? [WWW Document]. URL https://www.lucidchart.com/pages/swimlane-diagram?a=0 (accessed 3.4.19).

Lukas, J., Fridrich, J., Goljan, M., 2006. Digital camera identification from sensor pattern noise. IEEE Transactions on Information Forensics and Security 1, 205–214.

Marr, B., 2018. How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read [WWW Document]. Forbes. URL https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#63b33b5360ba (accessed 5.20.19).

Meikle, J., 2013. British Library adds billions of webpages and tweets to archive. The Guardian.

Meister, S., Chassanoff, A., 2014. Integrating Digital Forensics Techniques into Curatorial Tasks: A Case Study. International Journal of Digital Curation 9. https://doi.org/10.2218/ijdc.v9i2.325

Melão, N., Pidd, M., 2000. A conceptual framework for understanding business processes and business process modelling. Information systems journal 10, 105–129.

Microsoft, 2022. Products Ending Support for 2021 - Microsoft Lifecycle [WWW Document]. URL https://docs.microsoft.com/en-us/lifecycle/end-of-support/end-of-support-2021 (accessed 4.4.22).

Mitcham, J., Wheatley, P., 2019. Digital Preservation Coalition Rapid Assessment Model Ver 1. Digital Preservation Coalition. https://doi.org/10.7207/dpcram19-01

Moss, M., Gollins, T., 2017. Our digital legacy: an archival perspective. The Journal of Contemporary Archival Studies 4.

Muniswamy-Reddy, K.-K., Macko, P., Seltzer, M.I., 2010. Provenance for the Cloud., in: FAST. pp. 15–14.

Mvungi, J., Tossy, T., 2015. Usability evaluation methods and principles for the web. International Journal of Computer Science and Information Security 13, 86.

NAA, n.d. Managing social media | naa.gov.au [WWW Document]. URL https://www.naa.gov.au/information-management/types-information-and-systems/types-information/managing-social-media (accessed 4.4.22).

Narayana Samy, G., Shanmugam, B., Maarop, N., Magalingam, P., Perumal, S., Albakri, S.H., 2018. Digital Forensic Challenges in the Cloud Computing Environment, in: Saeed, F., Gazem, N., Patnaik, S., Saed Balaid, A.S., Mohammed, F. (Eds.), Recent Trends in Information and Communication Technology. Springer International Publishing, Cham, pp. 669–676.

NDSA Storage Infrastructure Survey Working Group, 2020. 2019 Storage Infrastructure Survey: Results of the Storage Infrastructure Survey. https://doi.org/10.17605/OSF.IO/UWSG7

NLA, 2021. Freedom of information disclosure log [WWW Document]. National Library of Australia. URL https://www.nla.gov.au/freedom-of-information-disclosure-log (accessed 2.1.22).

NLA, 2019. Trove - Australian Web Archive [WWW Document]. Trove. URL https://trove.nla.gov.au/help/categories/websites-category (accessed 2.8.21).

NLA, 2018. Privacy Policy [WWW Document]. URL https://www.nla.gov.au/policy-and-planning/privacy-policy (accessed 3.13.18).

NLA, 1999. Pandora Archive - Preserving and Accessing Networked Documentary Resources of Australia [WWW Document]. URL http://pandora.nla.gov.au/ (accessed 2.8.21).

Nohe, P., 2018. Re-Hashed: The Difference Between SHA-1, SHA-2 and SHA-256 Hash Algorithms. Hashed Out by The SSL Store$^{TM}$. URL https://www.thesslstore.com/blog/difference-sha-1-sha-2-sha-256-hash-algorithms/ (accessed 12.16.20).

NSLA, 2020. National and State Libraries Australia [WWW Document]. URL https://www.nsla.org.au/ (accessed 10.6.20).

NSLA, 2018. Changes at NSLA | National and State Libraries Australia [WWW Document]. URL https://www.nsla.org.au/news/changes-nsla (accessed 6.18.19).

NSLA, 2014. National position statement for Aboriginal and Torres Strait Islander library services and collections [WWW Document]. URL https://www.nsla.org.au/publication/national-position-statement-aboriginal-and-torres-strait-islander-library-services-and (accessed 6.6.18).

NSLA, 2013. Digital Preservation Policy 4th Edition (2013) [WWW Document]. URL https://www.nla.gov.au/policy-and-planning/digital-preservation-policy (accessed 11.28.19).

NSLA, 2010. Position statement on Indigenous intellectual property and ownership [WWW Document]. URL https://www.nsla.org.au/publication/position-statement-indigenous-intellectual-property-and-ownership (accessed 6.6.18).

OAIC, 2018a. Australian Privacy Principles guidelines [WWW Document]. Australian Government - Office of the Australian Information Commissioner. URL https://www.oaic.gov.au/agencies-and-organisations/app-guidelines/ (accessed 6.4.18).

OAIC, 2018b. Privacy business resource 21: Australian businesses and the EU General Data Protection Regulation [WWW Document]. Australian Government - Office of the Australian Information Commissioner. URL /agencies-and-organisations/business-resources/privacy-business-resource-21-australian-businesses-and-the-eu-general-data-protection-regulation (accessed 6.5.18).

OAIC, n.d. Rights and responsibilities [WWW Document]. Australian Government - Office of the Australian Information Commissioner. URL https://www.oaic.gov.au/privacy-law/rights-and-responsibilities (accessed 6.4.18).

OASIS, 2018. OASIS | Advancing open standards for the information society [WWW Document]. URL https://www.oasis-open.org/ (accessed 12.5.18).

OASIS, 2007. OASIS - WS-BPEL [WWW Document]. URL http://docs.oasis-open.org/wsbpel/2.0/OS/wsbpel-v2.0-OS.html (accessed 12.5.18).

Office of Parliamentary Counsel, C., 2017. Privacy Act 1988 Compilation No. 76 19, 27.

OGC, 2019. Keyhole Markup Language [WWW Document]. OGC Making location count. URL http://www.opengeospatial.org/standards/kml/

Olson, M., 2011. "Digital Forensics at Stanford University Libraries" Presentation at DPC briefing day - Digital Forensics for Preservation [WWW Document]. URL https://www.dpconline.org/docs/miscellaneous/events/628-forensics-olson/file

OMG, 2011. Business Process Model And Notation Specification Version 2.0 [WWW Document]. URL https://www.omg.org/spec/BPMN/2.0/ (accessed 12.3.18).

OPF, 2020. Digital preservation community survey results published. Open Preservation Foundation. URL https://openpreservation.org/news/digital-preservation-community-survey-results-published/ (accessed 10.2.20).

OPF, 2010. Format Identification for Digital Objects (fido) [WWW Document]. URL http://openpreservation.org/technology/products/fido/ (accessed 8.15.16).

Perdomo, E.G., Cardozo, M.T., Perdomo, C.C., Serrezuela, R.R., 2017. A Review of the User Based Web Design: Usability and Information Architecture. International Journal of Applied Engineering Research 12, 11685–11690.

Post, C., Chassanoff, A., Lee, C., Rabkin, A., Zhang, Y., Skinner, K., Meister, S., 2019. Digital Curation at Work: Modeling Workflows for Digital Archival Materials, in: 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL). pp. 39–48.

Queensland Ombudsman, 2017. What is a public interest disclosure? [WWW Document]. Queensland Ombudsman. URL https://www.ombudsman.qld.gov.au/improve-public-administration/public-interest-disclosures/what-is-a-public-interest-disclosure (accessed 3.14.18).

Quick, D., Choo, K.-K.R., 2016. Big forensic data reduction: digital forensic images and electronic evidence. Cluster Computing 19, 723–740. https://doi.org/10.1007/s10586-016-0553-1

Quick, D., Choo, K.-K.R., 2013. Forensic collection of cloud storage data: Does the act of collection result in changes to the data or its metadata? Digital Investigation 10, 266–277. https://doi.org/10.1016/j.diin.2013.07.001

Raghavan, S., 2013. Digital forensic research: current state of the art. CSI Transactions on ICT 1, 91–114.

Raghavan, S., Raghavan, S.V., 2014. Eliciting file relationships using metadata based associations for digital forensics. CSI Transactions on ICT 2, 49–64. https://doi.org/10.1007/s40012-014-0046-4

Rahman, N.H.A., Choo, K.-K.R., 2015. A survey of information security incident handling in the cloud. Computers & Security 49, 45–69. https://doi.org/10.1016/j.cose.2014.11.006

Redwine, G., Barnard, M., Donovan, K.M., Farr, E., Forstrom, M., Hansen, W.M., John, J.L., Kuhl, N., Shaw, S., Thomas, S.E., 2013. Born digital: Guidance for donors, dealers, and archival repositories. Council on Library and Information Resources Washington, DC.

Rivest, R., 1992. RFC1321: The MD5 message-digest algorithm.

Rolph, D., 2009. A critique of the national, uniform defamation laws. Torts Law Journal 16, 207–248.

Rosa, A., 2018. 10 website design do’s & don’ts [WWW Document]. Lucidpress Blog. URL https://www.lucidpress.com/blog/10-website-design-dos-donts (accessed 6.18.19).

Rousseau, Q., 2012. Thumbnail me. URL http://www.thumbnailme.com/ (accessed 12.5.18).

Roussev, V., McCulley, S., 2016. Forensic analysis of cloud-native artifacts. Digital Investigation 16, S104–S113. https://doi.org/10.1016/j.diin.2016.01.013

Rowell, C.J., Potvin, S., 2015. Preservation Metadata for Digital Forensics. A Report of the ALCTS PARS Preservation Metadata Interest Group Meeting. American Library Association Annual Meeting, Las Vegas, June 2014. Technical Services Quarterly 32, 320–325. https://doi.org/10.1080/07317131.2015.1031606

Sachowski, J., 2016. Chapter 1 - Understanding Digital Forensics, in: Sachowski, J. (Ed.), Implementing Digital Forensic Readiness. Syngress, Boston, pp. 3–16. https://doi.org/10.1016/B978-0-12-804454-4.00001-0

Schroffel, L., Soleau, T., Wang, L., 2018. 204.2 The evolution of digital preservation at the Getty Research Institute: How workflows have evolved in the past five years to address our varied needs. p. 5.

Shaw, S., 2017. DataAccessioner [WWW Document]. URL http://dataaccessioner.org/ (accessed 8.15.16).

SINTEF, 2013. Big Data, for better or worse: 90% of world's data generated over last two years -- ScienceDaily [WWW Document]. ScienceDaily. URL https://www.sciencedaily.com/releases/2013/05/130522085217.htm (accessed 5.20.19).

SLNSW, 2017. New Sensitive Collections Policy [WWW Document]. State Library of NSW. URL http://www.sl.nsw.gov.au/blogs/new-sensitive-collections-policy (accessed 6.4.18).

Software AG, 2020. Event-driven process chain (EPC) | ARIS BPM Community [WWW Document]. URL https://www.ariscommunity.com/event-driven-process-chain (accessed 12.3.18).

Spiceworks, 2019. The Future of Network and Endpoint Security [WWW Document]. Spiceworks. URL https://www.spiceworks.com/marketing/network-security/pdf-report/ (accessed 4.22.20).

State Archives & Records, 2015. Strategies for managing social media records [WWW Document]. URL https://www.records.nsw.gov.au/recordkeeping/advice/strategies-for-managing-social-media-information (accessed 4.4.22).

Tammaro, A.M., Matusiak, K., Sposito, F.A., Casarosa, V., Pervan, A., 2017. Understanding roles and responsibilities of data curators: an international perspective. Libellarium: journal for the research of writing, books, and cultural heritage institutions 9.

Tammaro, A.M., Matusiak, K.K., Sposito, F.A., Casarosa, V., 2019. Data curator's roles and responsibilities: an international perspective. Libri 69, 89–104.

The National Archives, 2020. UK Government Web Archive [WWW Document]. UK Government Web Archive. URL https://www.nationalarchives.gov.uk/webarchive/ (accessed 1.31.22).

The National Archives, 2018. Droid [WWW Document]. URL https://digital-preservation.github.io/droid/ (accessed 8.15.16).

Trehub, A., Davis, C., Jordan, M., May, C., Meister, S., 2018. LOCKSS Networks: Community-Based Digital Preservation. Digital Preservation in Libraries: Preparing for a Sustainable Future (An ALCTS Monograph).

Velte, A., Wikle, O.M., 2020. Scalable Born Digital Ingest Workflows for Limited Resources: A Case Study for First Steps in Digital Preservation. Preservation, Digital Technology & Culture (PDT&C) 49, 2–13.

Vincze, E.A., 2016. Challenges in digital forensics. Police Practice and Research 17, 183–194.

Vinh-Doyle, W.P., 2017. Appraising email (using digital forensics): techniques and challenges. Archives and Manuscripts 45, 18–30.

W3C, 2013. PROV Model Primer [WWW Document]. URL https://www.w3.org/TR/prov-primer/ (accessed 8.29.16).

Walters, T., Skinner, K., 2011. New roles for new times: Digital curation for preservation. Association of Research Libraries.

Wiedeman, G., 2016. Practical Digital Forensics at Accession for Born-Digital Institutional Records. The Code4Lib Journal.

Wilkinson, M.D., Dumontier, M., Aalbersberg, Ij.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., ’t Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B., 2016. The FAIR Guiding

Principles for scientific data management and stewardship. Scientific Data 3, 160018. https://doi.org/10.1038/sdata.2016.18

Willoughby, C., Frey, J.G., 2017. Documentation and visualisation of workflows for effective communication, collaboration and publication @ source. International Journal of Digital Curation 12, 72–87.

Wilsey, L., Skirvin, R., Chan, P., Edwards, G., 2013. Capturing and Processing Born-Digital Files in the STOP AIDS Project Records: A Case Study 4, 23.

Wiseman, C., 2016. A Practical Approach to Digital Preservation Planning at a Mid-Sized Academic Library, in: A Practical Approach to Digital Preservation Planning at a Mid-Sized Academic Library. Presented at the IFLA World Library and Information Congress, AUC Robert W. Woodruff Library Staff Publications. https://doi.org/10.22595/libpubs.00014

Woods, K., 2018. Understanding Bulk Extractor Scanners - BitCurator Environment - Confluence [WWW Document]. URL https://confluence.educopia.org/display/BC/Understanding+Bulk+Extractor+Scanners (accessed 10.29.18).

Woods, K., Lee, C.A., 2012. Acquisition and processing of disk images to further archival goals, in: Archiving Conference. Society for Imaging Science and Technology, pp. 147–152.

Zarour, K., Benmerzoug, D., Guermouche, N., Drira, K., 2019. A systematic literature review on BPMN extensions. Business Process Management Journal.

Zhao, J., Hartig, O., 2012. Towards Interoperable Provenance Publication on the Linked Data Web., in: LDOW.

Zhao, Z., Paschke, A., Zhang, R., 2016. A rule-based agent-oriented approach for supporting weakly-structured scientific workflows. Journal of Web Semantics 37–38, 36–52. https://doi.org/10.1016/j.websem.2016.02.002

# Appendix A – Tool Sources

**Archivematica -** https://www.archivematica.org/en/

**ArchivesSpace -** http://archivesspace.org/

**Archivists' Toolkit -** http://www.archiviststoolkit.org/

**Archon -** http://www.archon.org/

**Atom -** https://www.accesstomemory.org/en/

**Bagit/Bagger -** https://en.wikipedia.org/wiki/BagIt

**BitCurator -** https://bitcurator.net/

**Bulk Extractor -** https://www.forensicswiki.org/wiki/Bulk_extractor

**Catweasel -** http://www.softpres.org/glossary:catweasel

**Checksummer -** https://github.com/claudehohl/checksummer

**Cube-Tec CD-Inspector |**

**Cube-Tec Dobbin | -** https://www.cube-tec.com/en/solutions

**Cube-Tec Quadriga |**

**Curator's Workbench -** https://blogs.lib.unc.edu/cdr/index.php/about/cdr-development-and-collab/curators-workbench/

**DigiTool – no longer supported, see Rosetta**

**DROID -** http://www.nationalarchives.gov.uk/information-management/manage-information/policy-process/digital-continuity/file-profiling-tool-droid/

**Dublin Core -** http://dublincore.org/

**Encase -** https://www.guidancesoftware.com/encase-forensic

**Exiftool -** https://www.sno.phy.queensu.ca/~phil/exiftool/

**FC-5025 -** http://www.deviceside.com/fc5025.html

**Fedora -** https://duraspace.org/fedora/

**Ffmpeg -** https://www.ffmpeg.org/about.html

**FIDO -** http://fido.openpreservation.org/

**FITS -** https://projects.iq.harvard.edu/fits

**Fiwalk -** https://www.forensicswiki.org/wiki/Fiwalk

**Floppy Drive Controller -** https://en.wikipedia.org/wiki/Floppy-disk_controller

**FRED -** https://digitalintelligence.com/products/fred/

**FTK -** https://accessdata.com/products-services/forensic-toolkit-ftk

**FTK Imager – See FTK**

**Guymager -** https://guymager.sourceforge.io/

**Hydra – rebranded as Samvera** http://samvera.org/

**ICA-AtoM – no longer supported, see atom**

**ImageMagick -** https://imagemagick.org/index.php

**JHOVE -** http://jhove.openpreservation.org/

**Kryoflux -** https://www.kryoflux.com/?page=kf_features

**LibreOffice -** https://www.libreoffice.org/discover/libreoffice/

**LOCKSS -** https://www.lockss.org/about/what-is-lockss/

**Mediainfo -** https://mediaarea.net/en/MediaInfo

**MetaArchive -** https://metaarchive.org/

**METS -** http://www.loc.gov/standards/mets/

**NZME -** http://meta-extractor.sourceforge.net/

**OAIS -** https://en.wikipedia.org/wiki/Open_Archival_Information_System

**OSFMount -** https://www.osforensics.com/tools/mount-disk-images.html

**oXygen -** https://www.oxygen-forensic.com/en/

**PREMIS -** https://www.loc.gov/standards/premis/

**Robocopy -** https://en.wikipedia.org/wiki/Robocopy

**Rosetta -** https://www.exlibrisgroup.com/products/rosetta-digital-asset-management-and-preservation/

**Rsync -** https://en.wikipedia.org/wiki/Rsync

**Steinberg Wavelab v9 -** https://www.steinberg.net/en/products/wavelab/start.html

**Tableau -** https://www.tableau.com/

**Tiffinfo -** http://www.libtiff.org/tools.html

**veraPDF -** https://verapdf.org/home/

**Wiebetech Cru Write Blocker -** https://www.cru-inc.com/products/wiebetech/

# Appendix B - Questionnaire

Thank you for taking the time to participate in the following questionnaire. The purpose is to better understand the digital preservation process your institution follows and how/where forensic techniques are being used.

Your response will be used for research into the design of more effective and efficient workflows that better utilise digital forensics to aid preservation through all stages.

Please answer as much as you can in however many words you wish within the spaces allocated between each question. Please feel free to user hyperlinks or attach any supporting documentation in the email when submitting.

**Please note, the institution and the participant of this questionnaire will not be identified in any way and will remain completely anonymous.**

---

**Donor Agreements / Ethical standards**

**Question 1:** Once material is submitted by a donor and then accepted, what is the process that follows to determine ownership, access, and donor stipulations?

**Question 2:** When dealing with donated digital material, what is the process when sensitive information or content is discovered that **<u>has</u>** been addressed in the donor agreement?

**Question 3:** Regarding question **2**, what process follows if something is discovered that **<u>has not</u>** been addressed in the donor agreement?

**Question 4:** If the donor or the next of kin are no longer available and sensitive information or content is discovered, what is the standard ethical procedure when dealing with such data?

***Example*** – You are processing data on a person/group that is to be made publicly available and sensitive content is discovered that would be detrimental to that person's or group's reputation. Do you make that data public or redact it? Furthermore, do factors such as, `Is the person or group still alive' or `Is the subject politically based?' change the protocol?

**Digital Preservation**

**Question 1:** What are the common types of born-digital content your institution works with? (File types, documents, Image, video, audio, etc.)

**Question 2:** When preserving digital content, are there processes involved to add additional metadata (descriptive metadata) to give the digital content context, as well as improving search and retrieval functionality? (this does not include environment or dependency description)

**Question 3:** Please describe the process and list any tools (hardware/software) used for this process.

**Question 4:** What precautions are in place to ensure digital content is not changed or accidentally modified during ingest and through to storage?

**Question 5:** What software is used to facilitate the preservation process? (name and version, please)

**Question 6:** What is the purpose of the specified software? (E.g. which part of the process does the software facilitate or does it have a unique function?)

**Question 7:** Is there a workflow model diagram of the digital preservation process available and are you willing to share it?

If a diagram is not available, are you able to list the steps in a typical preservation process from ingest to storage, including maintenance? Please specify where human intervention is needed and when a process is automated.

**Question 8:** Is there part of your workflow that you believe would be made more efficient by developing specific software? This could be either because there is no software available or the currently available software does not suit your purpose.

**Digital Forensics**

**Question 1:** Please list any forensic hardware and software used: (Primarily forensic software that is typically used for forensic analysis/criminology, but repurposed for born-digital preservation)

**Question 2:** If existing forensic software could be used for something other than its typical function and the accompanying documentation was amended to accommodate digital preservation, would this be a step towards adopting new and possibly better methods?

**Question 3:** If there are budget concerns and expensive proprietary software is out of the question, are open-source, freely available tools considered and or accepted?

**Question 4:** If there are solutions to improve current working procedures, would the library/archive be open to reviewing suggested theoretical workflow improvements and amendments?

## Appendix C – Tool Data

Table 6 - BitCurator Consortium - 2012 Dataset

| Tool | Total | MEM1 | MEM2 | MEM3 | MEM4 | MEM5 | MEM6 | MEM7 |
|---|---|---|---|---|---|---|---|---|
| **Archivists' Toolkit** | 3 | x | x | | | | x | |
| **Archivematica** | 1 | | | | | | | x |
| **Atom** | 1 | | | | | | | x |
| **Bagit/Bagger** | 3 | x | | | | | x | x |
| **Catweasel** | 2 | x | | | | | x | |
| **Curator's Workbench** | 1 | | | | x | | | |
| **dd (unix utility)** | 2 | | | x | | | | x |
| **Encase** | 1 | | | | | | | x |
| **Excel** | 2 | | | x | x | | | |
| **FC-5025** | 2 | x | | | | | x | |
| **Fedora** | 1 | | | | | x | | |
| **FIDO** | 2 | x | | | | | x | |
| **FITS** | 1 | | | | | | | x |
| **Fiwalk** | 2 | x | | | | | x | |
| **FRED** | 1 | | | | | x | | |
| **FTK** | 3 | x | | | | x | x | |
| **FTK Imager** | 4 | x | x | | x | | x | |
| **ICA-AtoM** | 1 | | | | | | | x |
| **JHOVE** | 1 | x | | | | | | |
| **Kryoflux** | 3 | x | x | | | | x | |

Table 7 - BitCurator Consortium – 2016 Dataset

| Tool | Total | MEM8 | MEM9 | MEM10 | MEM11 | MEM12 |
|---|---|---|---|---|---|---|
| **Archivists' Toolkit** | 1 | | | x | | |
| **ArchivesSpace** | 2 | x | x | | | |
| **Archon** | 1 | | | | x | |
| **Bagit/Bagger** | 3 | | | x | x | x |
| **BitCurator** | 5 | x | x | x | x | x |
| **Bulk Extractor** | 3 | | x | x | | x |
| **DROID** | 2 | | | x | x | |
| **Encase** | 1 | | | | x | |
| **Excel** | 1 | | | | | x |
| **FC-5025** | 2 | | x | | | x |
| **Fedora** | 1 | x | | | | |
| **FITS** | 1 | | x | | | |
| **Fiwalk** | 4 | | x | x | x | x |
| **FRED** | 3 | x | | x | x | |
| **FTK Imager** | 2 | | x | | x | |
| **Guymager** | 2 | | x | x | | |
| **hydra** | 1 | | x | | | |
| **Kryoflux** | 2 | | | | x | x |
| **LibreOffice** | 1 | | | x | | |
| **MetaArchive** | 1 | | | | x | |
| **NZME** | 1 | | | | x | |
| **oXygen** | 1 | | | x | | |
| **Robocopy** | 1 | | | x | | |
| **Rsync** | 1 | | | x | | |

“x” Indicates the tool was used within BitCurator.

Table 8 - Australian Dataset 2018

| Tool | Total | AU1 | AU2 | AU3 | AU4 | AU5 | AU6 | AU7 |
|---|---|---|---|---|---|---|---|---|
| Acrobat Pro XI | 1 | x | | | | | | |
| Adobe Premier Pro | 1 | | x | | | | | |
| Archivematica | 1 | | | | | | x | |
| Bagit/Bagger | 2 | | x | x | | | | |
| BitCurator | 2 | x | x | | | | | |
| Checksums (unknown) | 1 | | | | | x | | |
| Checksummer | 1 | | | | x | | | |
| CSV | 1 | | | x | | | | |
| Cube-Tec CD-Inspector | 1 | x | | | | | | |
| Cube-Tec Dobbin | 1 | x | | | | | | |
| Cube-Tec Quadriga | 1 | x | | | | | | |
| DigiTool | 1 | | | | | x | | |
| DROID | 1 | | | | x | | | |
| Exiftool | 1 | | | x | | | | |
| ffmpeg | 1 | | | | | x | | |
| FRED | 2 | x | | | x | | | |
| FTK | 2 | | | x | x | | | |
| FTK Imager | 3 | x | x | x | | | | |
| HeX Editors | 1 | | | | | x | | |
| LMS | 1 | | | | x | | | |
| LOCKSS | 1 | | x | | | | | |
| mediainfo | 1 | | | | | x | | |
| METS | 1 | | | x | | | | |
| OSFMount | 1 | x | | | | | | |
| Plextor | 1 | x | | | | | | |
| Rosetta | 2 | | | x | | x | | |
| Steinberg Wavelab v9 | 1 | x | | | | | | |
| Floppy Drive Controller | 1 | | x | | | | | |
| Tableau | 1 | x | | | | | | |
| Tiffinfo | 1 | | | | | x | | |
| Catalogue (Unique) | 1 | | | | | x | | |
| Wiebetech Cru Write Blocker | 1 | | x | | | | | |
| Write Blocker (unknown) | 2 | | | x | x | | | |

# Appendix D – Workflow Evaluation

Table 9 - BitCurator Consortium - Workflow Criteria

| Workflow | Donor Agreement | Sensitive Data Discovery | Sensitive Data Handling |
|---|---|---|---|
| **MEM1** | x | | |
| **MEM2** | x | x | x |
| **MEM3** | x | | |
| **MEM4** | x | x | x |
| **MEM5** | x | x | x |
| **MEM6** | x | | |
| **MEM7** | x | | |
| **MEM8** | x | x | x |
| **MEM9** | | x | |
| **MEM10** | x | x | |
| **MEM11** | x | x | x |
| **MEM12** | x | x | |
| | | | |
| | **Donor Agreement** | **Sensitive Data Discovery** | **Sensitive Data Handling** |
| **Total** | 11 | 8 | 5 |
| **Uncertain** | 1 | 1 | 3 |

Table 10 - OSSArcFlow - Workflow Criteria

| Workflow | Donor Agreement | Sensitive Data Discovery | Sensitive Data Handling |
|---|---|---|---|
| OSS1 | | | |
| OSS2 | x | | |
| OSS3 | x | x | x |
| OSS4 | x | x | x |
| OSS5 | x | | |
| OSS6 | x | x | |
| OSS7 | | | |
| OSS8 | x | | |
| OSS9 | x | x | |
| OSS10 | x | x | x |
| OSS11 | x | x | x |
| OSS12 | x | x | x |
| | | | |
| | Donor Agreement | Sensitive Data Discovery | Sensitive Data Handling |
| Total | 10 | 7 | 5 |
| Uncertain | 0 | 0 | 0 |

# Appendix E – Publications Resulting from this Thesis

**Title:** Metadata Provenance and Vulnerability

**Authors:** Timothy Robert Hart, Denise de Vries

**Publication:** Information Technology and Libraries, Vol 36, Issue 4, P 24-33, 2017

**Abstract:**

The preservation of digital objects has become an urgent task in recent years as it has been realised that digital media have a short life span. The pace of technological change makes accessing these media increasingly difficult. Digital preservation is primarily accomplished by main methods, migration and emulation. Migration has been proven to be a lossy method for many types of digital objects. Emulation is much more complex; however, it allows preserved digital objects to be rendered in their original format, which is especially important for complex types such as those comprising multiple dynamic files. Both methods rely on good metadata to maintain change history or construct an accurate representation of the required system environment. In this paper, we present our findings that show the vulnerability of metadata and how easily they can be lost and corrupted by everyday use. Furthermore, this paper aspires to raise awareness and to emphasise the necessity of caution and expertise when handling digital data by highlighting the importance of provenance metadata.

**Title:** Australian Law Implications on Digital Preservation

**Authors:** Timothy Robert Hart, Denise de Vries, Carl Mooney

**Publication:** Accepted in iPres 2019, Amsterdam, Sep. 16, 2019.

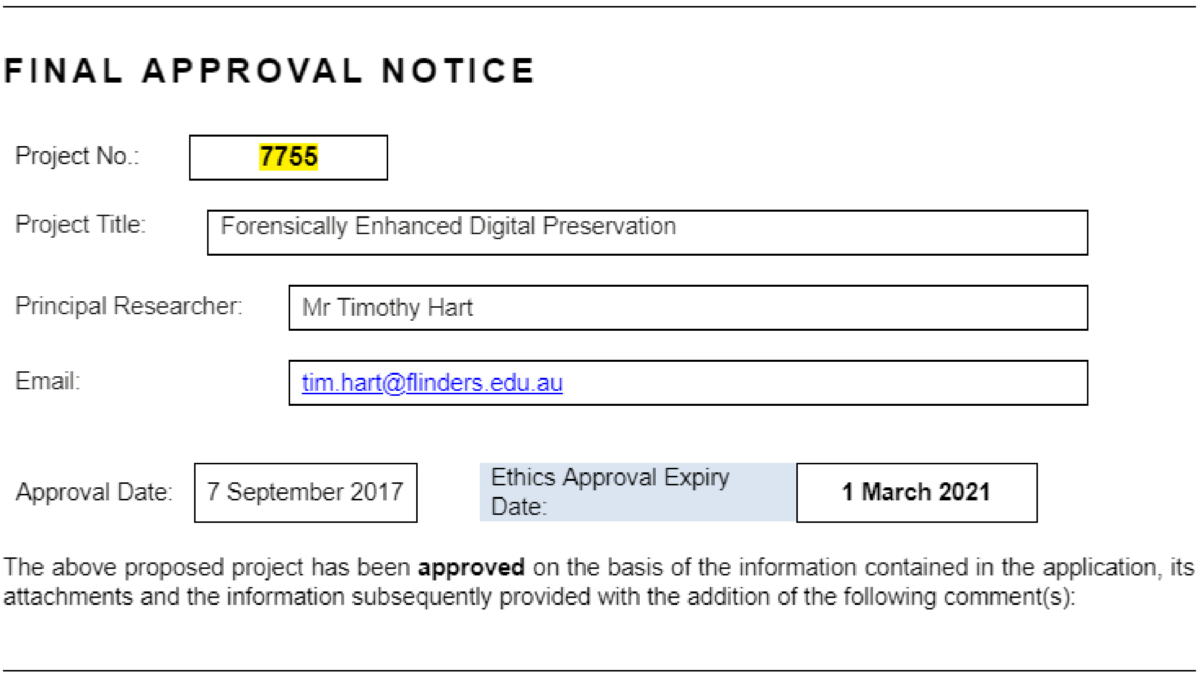
DOI 10.17605/OSF.IO/EZ6FQ

https://osf.io/ez6fq/

**Abstract:**

Collection institutions (Libraries, Archives, Galleries, and Museums) are responsible for storing and preserving large amounts of digital data, which can range from historical/public figure records, to state or countrywide events. The ingest process often requires sifting through large amounts of data which may not always be sorted or categorised from the source/donor. It is possible to discover information that was not intended to be disclosed should the donor not be privy to the existence of said material. This issue is typically handled by communicating with the donor; however, if they have no relation to what has been uncovered in the data, further steps may need to be taken. If the data belongs to or is about someone living, that person may need to be contacted, depending on the nature of the data discovered. If the person of interest is no longer living, legally there would no issue disclosing all information uncovered. Implications on living relatives must be considered should the disclosed information be potentially revealing or harmful to them. This can include hereditary health issues, political or religious views, and other sensitive information. There are significantly more variables to consider, such as public interest and defamation which can heavily impact the decision process following the discovery of sensitive data, all whilst guided, but not necessarily enforced by Australian law. This remains somewhat of a grey area as the entities handling such data are often exempt from these laws and principles, making these decisions ethically and morally based more so than legally. The laws and policies that surround privacy issues, defamation, and data relating to Aboriginal and Torres Strait Islander people and culture are explored. The aim is to raise awareness on potential issues that may arise in collection institutions as well as potential threats already sitting in storage and the laws and policies that may serve as guidelines to help overcome/mitigate such issues.

# Appendix F – Ethics Documentation



## FINAL APPROVAL NOTICE

| | |
|---|---|
| Project No.: | **7755** |
| Project Title: | Forensically Enhanced Digital Preservation |
| Principal Researcher: | Mr Timothy Hart |
| Email: | tim.hart@flinders.edu.au |

| | | | |
|---|---|---|---|
| Approval Date: | 7 September 2017 | Ethics Approval Expiry Date: | **1 March 2021** |

The above proposed project has been **approved** on the basis of the information contained in the application, its attachments and the information subsequently provided with the addition of the following comment(s):

# INFORMATION SHEET

## (for Library, Archive, Galleries, and Museum Staff)

**Title:** Forensically Enhanced Digital Preservation

**Researchers:**

Mr Tim Hart

College of Science and Engineering

Flinders University

Ph: +61 8 8201 3639

**Supervisor(s):**

Dr Denise de Vries

College of Science and Engineering

Flinders University

Ph: +61 8 8201 3639

**Description of the study:**

This study is part of the project entitled Forensically Enhanced Digital Preservation. This project will investigate how to improve current digital preservation workflows in libraries, archives, galleries, and museums through potentially unknown forensic tools and methods. This project is supported by Flinders University College of Science and Engineering department.

**Purpose of the study:**

This project aims to:

- increase awareness of tools and methods that could be used for digital preservation, but are not necessarily designed for that purpose
- design more effective and efficient workflows
- amend manuals and supporting documentation for specific forensic software to accommodate a digital preservation perspective

**What will I be asked to do?**

You are invited to participate in a questionnaire containing questions about your institutions workflow process to establish an idea of what methods and techniques are being used in local libraries, archives, galleries, and museums. about. Participation is entirely voluntary. The questionnaire will take about 30 minutes to 2 hours. Once returned, the data will be collected for analytics, stripping it from identifying factors.

### What benefit will I gain from being involved in this study?

By participating in this study, you will help in identifying accurately how the current establishments mentioned above are handling digital preservation. With this information, better and easier solutions may be made available, concluding to an improved workflow that your institution may be able to implement.

### Will I be identifiable by being involved in this study?

The only identifiable information will be linking the institution with the established workflow derived from the questionnaire if adequate information is supplied. You may however choose to have your institution remain anonymous and it will not impact the study. You as the participant are not identifiable in any way. The data published from this study will only output software and processes that make up a digital preservation workflow.

### Are there any risks or discomforts if I am involved?

There are no risks involved nor will there be any discomfort. If you have any concerns regarding anticipated or actual risks or discomforts, please raise them with the investigator.

### How do I agree to participate?

Participation is voluntary. You may answer 'no comment' or refuse to answer any questions and you are free to withdraw from the project at any time without effect or consequences. A consent form accompanies this information sheet. If you agree to participate please read and sign the form and send it to tim.hart@flinders.edu.au. The questionnaire has also been attached for you to review before deciding.

### How will I receive feedback?

Outcomes from the project will be summarised and used within the final report of the project. The report will be made available once completed and published. Should you have any further questions or wish to request specific feedback, please contact the investigator on the email provided above.

Thank you for taking the time to read this information sheet and we hope that you will accept our invitation to be involved.

*This research project has been approved by the Flinders University Social and Behavioural Research Ethics Committee (Project number 7755.). For more information regarding ethical approval of the project the Executive Officer of the Committee can be contacted by telephone on 8201 3116, by fax on 8201 2035 or by email human.researchethics@flinders.edu.au*

# LETTER OF INTRODUCTION

This letter is to introduce Mr Tim Hart who is a post-graduate student in the College of Science and Engineering, Flinders University.

He is undertaking research leading to the production of a thesis or other publications on the subject of digital forensics and digital preservation

He would like to invite you to assist with this project completing a questionnaire which covers certain aspects of this topic. No more than 1-2 hours on one occasion would be required.

Be assured that any information provided will be treated in the strictest confidence and none of the participants will be individually identifiable in the resulting thesis, report or other publications. You are, of course, entirely free to discontinue your participation at any time or to decline to answer particular questions.

Any enquiries you may have concerning this project should be directed to me at the address given above or by telephone on 08 8201 3639 fax 08 8201 2904 or e-mail denise.devries@flinders.edu.au

Thank you for your attention and assistance.

Yours sincerely

DdeVries

**Lecturer**

**Computer Archaeology Laboratory**<br>
**College of Science & Engineering**<br>
**Flinders University**

*This research project has been approved by the Flinders University Social and Behavioural Research Ethics Committee (Project number 7755). For more information regarding ethical approval of the project the Executive Officer of the Committee can be contacted by telephone on 8201 3116, by fax on 8201 2035 or by email human.researchethics@flinders.edu.au*