



Novice to Know-How: Email Preservation

Complete Learning Pathway Text



Digital**Preservation**Coalition

| | |
|----------|--|
| THE | |
| NATIONAL | |
| ARCHIVES | |

Contents

| | |
|---|----|
| Introduction..... | 1 |
| Key Learning Objectives..... | 1 |
| Acknowledgements | 1 |
| License Information | 1 |
| Course 1: Introduction to Email Preservation | 2 |
| Module 1.1: Why Preserve Email? | 2 |
| Module 1.2: Key Issues in Email Preservation | 3 |
| Course 2: How Email Works | 7 |
| Module 2.1: How Email is Sent..... | 7 |
| Module 2.2: What's in An Email..... | 8 |
| Module 2.3: More on Email Standards | 11 |
| Module 2.4: Email Clients | 14 |
| Module 2.5: Email Preservation Formats | 15 |
| Course 3: Developing an Email Preservation Program..... | 19 |
| Module 3.1: Understanding How Email is Managed | 19 |
| Module 3.2: Advocating for Email Preservation | 19 |
| Module 3.3: Developing Policy..... | 22 |
| Module 3.4: Promoting Good Practice for Email Management | 24 |
| Module 3.5: Legal Contexts for Email Preservation..... | 26 |
| Module 3.6: Retention and Deletion | 30 |
| Module 3.7: Thinking About Email Preservation Workflows | 32 |
| Module 3.8: Introduction to Email Preservation Tools and Systems | 35 |
| Course 4: Selection and Capture..... | 37 |
| Module 4.1: Selection Methods | 37 |
| Module 4.2: Focus on the Capstone Approach | 40 |
| Module 4.3: Capturing Email Archives..... | 42 |
| Module 4.4: Capturing Email from Outlook..... | 45 |
| Module 4.5: Capturing Email from Gmail..... | 48 |
| Module 4.6: Introduction to ePADD, Including Capture/Import of Email..... | 51 |
| Course 5: Appraisal and Processing | 57 |

| | |
|---|-----|
| Module 5.1: Processing Email for Ingest | 57 |
| Module 5.2: Appraisal Decisions | 60 |
| Module 5.3: The ePADD Appraisal Module | 62 |
| Module 5.4: Using the ePADD Appraisal Module..... | 70 |
| Module 5.5: Capturing Metadata..... | 78 |
| Course 6: Preservation..... | 81 |
| Module 6.1: Preservation Methods and Email | 81 |
| Module 6.2: Designing an Archival Information Package for Email | 83 |
| Module 6.3: The ePADD Processing Module | 86 |
| Course 7: Discovery and Access | 95 |
| Module 7.1: Facilitating Discovery for Preserved Email | 95 |
| Module 7.2: Providing Access | 96 |
| Module 7.3: The ePADD Discovery and Delivery Modules | 98 |
| Course 8: Email Preservation Case Studies | 105 |
| Module 8.1: UK Cabinet Office Case Study | 105 |
| Module 8.2: Trinity College Dublin Case Study..... | 111 |
| Module 8.3: Sheffield City Council Archives Case Study | 118 |

Introduction

This document provides a complete text surrogate of the content included in the Novice to Know-How (N2KH) Email Preservation learning pathway. The learning pathway aims to provide learners with a comprehensive and practical introduction to email preservation. Upon completing the course, learners should understand the key organizational issues relating to email preservation and feel empowered to develop workflows to facilitate the preservation of email.

Key Learning Objectives

After completing this learning pathway, learners will be able to:

1. Explain what email preservation is and why it is important
2. Plan and implement an email preservation program at their organization
3. Develop and execute email preservation workflows

Acknowledgements

The creation of the N2KH Email Preservation learning pathway was conceived of and funded by The National Archives (UK) as part of their [“Plugged In, Powered Up”](#) digital capacity building strategy for the UK Archives Sector. The learning pathway content and design was developed by the Digital Preservation Coalition (DPC) in consultation with The National Archives (UK). Many thanks to colleagues from the DPC and The National Archives (UK) who help review and refine the content of this learning pathway.

This learning pathway is offered in memory of Dr. Jo Pugh, The National Archives (UK), a devoted advocate for building digital skills within the UK Archives and Cultural Heritage Sectors, and without whom Novice to Know-How would not exist.

License Information

The N2KH pathway is made available for reuse under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International \(CC BY-NC-SA 4.0\)](#) license. This allows for free reuse of the content for non-commercial purposes with attribution (as below), and any reworked content that is shared to 3rd parties must carry the same license.

The following attribution statement should be prominently displayed alongside the N2KH training content, at a minimum once per course through the learning pathway.

This learning pathway was developed by the Digital Preservation Coalition, in collaboration with The National Archives (UK). It was conceived of and funded by The National Archives (UK) as part of the "Plugged In, Powered Up" digital capacity building strategy.

Course 1: Introduction to Email Preservation

Module 1.1: Why Preserve Email?

Email is one of the most pervasive communication methods of the digital age. Despite some claims that the demise of email is imminent, it remains a key tool for sharing information in both personal and professional spheres. It encompasses both formal and informal communications and can include information on almost any topic. While preserving email is slowly becoming a more common element of digital preservation programs, it is still far from being routine. In this module we will examine some of the reasons why it is important to take proactive steps to preserve email.

In his “Preserving Email” Technology Watch Report, Chris Prom describes the essential role email plays in our work and personal lives, referring to it as an “evidence rich trail of actions, thoughts, and communications”. He also observes that in recent years email has played a central role in high profile legal cases, in news exposés, and in the political world. As an example of its influence, we can look at the effect email potentially had on the outcome of the 2016 US Presidential Election, where an email-centric controversy significantly impacted Hillary Clinton’s campaign.

Email within an organizational setting can contain records of decisions made, of the development of strategy and plans, of meetings held, of documents shared, of actions performed, and much more. It cuts across the full range of our workplace activities in a way that is unique among the types of records we preserve.

This is also true of email used in our personal lives. A mailbox can include messages that touch on a wide range of topics including a person’s finances, health, social interactions, purchases, services used, entertainment, hobbies, and family. Preserving the email account of an individual can provide an excellent resource of information about how they lived their life.

Due to the wide range of issues covered in email messages there are often many legal reasons to retain and preserve them. For an organization this might be to meet regulatory requirements or institutional retention policies, for an individual perhaps as evidence of a financial transaction where a claim might need to be made. As mentioned earlier, email plays an increasingly significant role in many legal cases and a robust approach to email preservation will help to maintain its evidential value.

Additionally, it is not just the content of email that makes it an important record format, but also the qualities that are inherent in email technology itself. The ease of access to and use of email, the speed of communication thanks to its almost instantaneous delivery, and the connections and context built from threads of replies add an additional layer of information and meaning that can be parsed from preserved email. The “Future of Email Archives” report from the Task Force on Technical Approaches for Email Archives describes email as a “story keeper”, stating that “email doesn’t just document digital life; it documents life itself”. This is

perhaps one of the most important reasons to preserve email, not just as a simple record but as a microcosm of our work or personal lives.

As with all digital preservation, before embarking on an email preservation journey it is useful to consider why you wish to preserve this digital content. A clear understanding of the “why” will help with decision making as you build your email preservation program. It can inform all aspects of the work, including policy decisions, retention planning, tool choices, workflow development, and the provision of access.

Now take a few minutes to consider why your organization wishes to preserve email. Note down a few reasons before moving on.

Now we have identified some of the reasons for preserving email. The next module will introduce key issues you will need to consider when establishing an email preservation program.

Module 1.2: Key Issues in Email Preservation

Introduction

The preservation of email shares much in common with the preservation of other types of digital content, but it also presents a number of unique challenges. For example, emails aren't just individual documents but interconnected messages that can form conversations between multiple authors. These interconnections can provide important information in themselves and may need to be preserved along with the content contained within individual emails. In this module we will provide an overview of those challenges, which will be examined in more detail as you progress through this learning pathway.

Email is Disposable

Discussions of preserving digital content often begin with the difficulties of working with content that is by its very nature ephemeral. There is no definitive object in the digital world, many copies may exist in different locations, and they may be easily deleted or lost. Capturing content requires proactive identification of what is worthy of preservation and collecting a copy as soon as possible. While this is a concern for all content types, it is one of, if not the, biggest challenges in email preservation.

As a medium, email is often viewed by users as fleeting and disposable. Who hasn't deleted an email without giving the process much thought? Therefore, we are immediately starting from a position where the survival of emails might be reliant on the good will of the user and their understanding of the value of email as a record. This is often further compounded by the organizational contexts in which email and its users exist.

Organizational Issues

For many organizations, non-current emails are seen as more of a liability than an asset, something to be disposed of as soon as is convenient to avoid risk. This may even be embedded into policy on the use of email, encouraging users to delete messages that they see no further use for. Removal of email is often also part of an organization's information technology practices. Older emails may be deleted to free up storage space or they might be moved to an "archive" storage area (it is important to note that the term "archive" is used differently in information technology than it is in information management). Some organizations even have automated deletion of emails that have not been identified as records once they reach a certain age, for example a year after they were sent.

Attempts to address these issues have included encouraging users to log important emails in an organization's document and records management system. This approach has generally been met with mixed levels of success. There is often still confusion about which emails should be included as records, while some employees resent the additional work involved, particularly the required metadata creation, and will often skip the process.

Ultimately, combating these issues requires a combination of activities including advocacy, policy updates, awareness raising of and training in good email management practice, and collaboration with colleagues from information technology teams. This will hopefully lead to the inclusion of email retention and preservation within the organization's record-keeping practices.

Quantity of Email

Challenges also arise from the sheer quantity of email that is generated and how it is then organized or "filed" by the user. Even in a single email account, the number of emails contained might run into the tens of thousands, particularly if the member of staff has been in post for a significant period of time. And, unless an organization has strict guidance on the filing of emails, the emails will be arranged at the will of the user. This could be by project, function, correspondent, date, or one of many other schemes. The quantity and arrangement of emails are issues that will impact on selection and appraisal decisions and processes.

Policy and Legal Contexts

Next, we must consider the policy and legal contexts that email exists within. Depending on the business conducted by an organization, there may be legal and regulatory requirements for retention of records that will impact email. It is important that these are incorporated into both policy and processes for email management and preservation, as should consideration of emails with potential historical or research value. Personal or sensitive information is also often shared within emails and so preservation approaches must consider data protection and freedom of information requirements.

The contents of an email mailbox can also present complex intellectual property rights (IPR) issues and considerations relating to the ethics of collecting. The emails received will have been created by a variety of different authors who will likely not have provided consent for their content to be collected, preserved, and shared with others. These issues may then be further compounded by the inclusion of attachments that may have been created by a third party rather than the email author. These IPR and ethical issues will need to be considered as part of any email preservation work.

Technical Challenges

There are also a number of technological challenges that are unique to email preservation. One key issue relates to the implementation of the various standards that govern the structure of emails and their transmission. While these standards allow for both consistency and flexibility in the transmission of emails, they are open to different interpretations and implementations in the software, known as email clients, that users operate to access and manage their email.

The inherent complexity of email as a format also causes preservation challenges. In addition to the information contained within the body of an email, important context can also be derived from the relationships between emails, for example from the threads of a continuing conversation, and we must consider how to retain these links between messages. Email systems are also now often integrated with other productivity tools such as messaging platforms like Slack, work planning apps, or customer relationship management systems. We need to consider if we need to maintain these links during the preservation process.

Additionally, while the basic structure of an email is relatively well standardized, there are no such constraints on the attachments that can be added beyond the size limits different email systems enforce. Attachments can contain files of any format, including container formats such as ZIP which in turn can contain a multitude of formats within the zipped folder. Emails often also contain links to external resources, such as websites. We, therefore, must also consider how to preserve these attachments and links. Do we maintain them alongside the original emails, or do we preserve them separately and record the connection?

Discovery and Access

Finally, we must consider the discovery and access requirements that are specific to email. How do we catalog email collections to make them findable? Do we provide access to a complete email inbox or individual emails? How do we manage access to collections that contain personal or other sensitive data? Can we provide online access or will access need to be provided onsite only? Access provision will be affected by a combination of preservation decisions that have been made and the resources that are available.

Wrap-Up

This module has introduced some of the key issues that must be considered when preserving email. Each of these issues will be examined in more detail as you work through this learning pathway. The next set of modules will delve into the standards, structures, and processes of email, as an understanding of these will aid you in establishing your email preservation program and inform decisions you will need to make.

Course 2: How Email Works

Module 2.1: How Email is Sent

For most of us, the process of sending an email is straightforward and routine. We specify a recipient, add a subject, compose a message, and hit send. But what happens next? Understanding the basics of how email works will help us make and implement preservation decisions. In this module, we'll look at what happens when you hit send and how your message is transferred from your mail account to its recipients. A diagram is included below to provide a visual representation of what is described in this module.

Sometimes sending an email seems like magic, you hit send and, usually, the message pops up in the recipient's mailbox within seconds, no matter where in the world they are located. But to make this happen there is a global system employing a combination of protocols, servers, routers, and agents working behind the scenes.

The journey of any email begins with the software or webmail service you use to access your email account, such as Outlook, Gmail, or Apple Mail. This is known as the email client, or more formally as the Mail User Agent (MUA). The Mail User Agent assembles the email information you have entered into a standardized format and then packages it into an "envelope" that has the key pieces of information required by email transfer protocols. These include your email address and the email address of the recipient.

The envelope containing the email is then passed to your mail service's Mail Transfer Agent (MTA), also often known as an email or exchange server. The Mail Transfer Agent checks to see if the email is to be delivered locally, to a recipient using the same server, if so it will deliver the email to the mailbox of the correct user. If the email is not being delivered locally, it is added to the outgoing email queue.

When the email clears the outgoing queue it enters the Internet's network cloud where it is routed along a chain of connected servers until it reaches the correct server for the recipient. The route the email takes is generally not predetermined, relying instead on server availability. For example, an email sent in the UK to another recipient in the UK might pass through servers anywhere in the world while being transferred. This makes email transfer fast and flexible, but is also the reason it is susceptible to spam and hacking.

To determine where an email should be delivered, Mail Transfer Agents make use of the Internet's Domain Name System (DNS) to find the correct server. This process starts by identifying the correct top-level domain from the recipient email address, such as .com, .org, or.co.uk. Then the Mail Transfer Agent will access the root servers for this domain to find which mail exchange (MX) server accepts emails for the subdomain of the address. These checks may happen several times as the email moves from server to server before it arrives at the correct one.

When the email reaches the correct server, the Mail Transfer Agent on the final server asks if the host server accepts messages for the username included in the email address and, if it does, transfers the email to the server for the recipient's domain. At this stage the email will likely also encounter spam filters, virus scans, firewalls, and other hurdles, such as email size limits, established by the server's admins to ensure their systems wishes to accept the email. If the email doesn't pass these checks it will be returned to the sender.

Once the email has been accepted, the server's local Mail Transfer Agent will pass the message to a Mail Delivery Agent (MDA) which will deliver it to the correct mailbox. At this point the envelope is normally removed and discarded. The email will then be held on the server until the recipient logs in and uses their email client software or service to access and open it.

There are several standards that are used during this process to ensure the interoperability of emails between servers and clients, and to facilitate their transfer across the Internet. We will introduce some of these standards in our next module where we'll look at the structure of an email message itself.

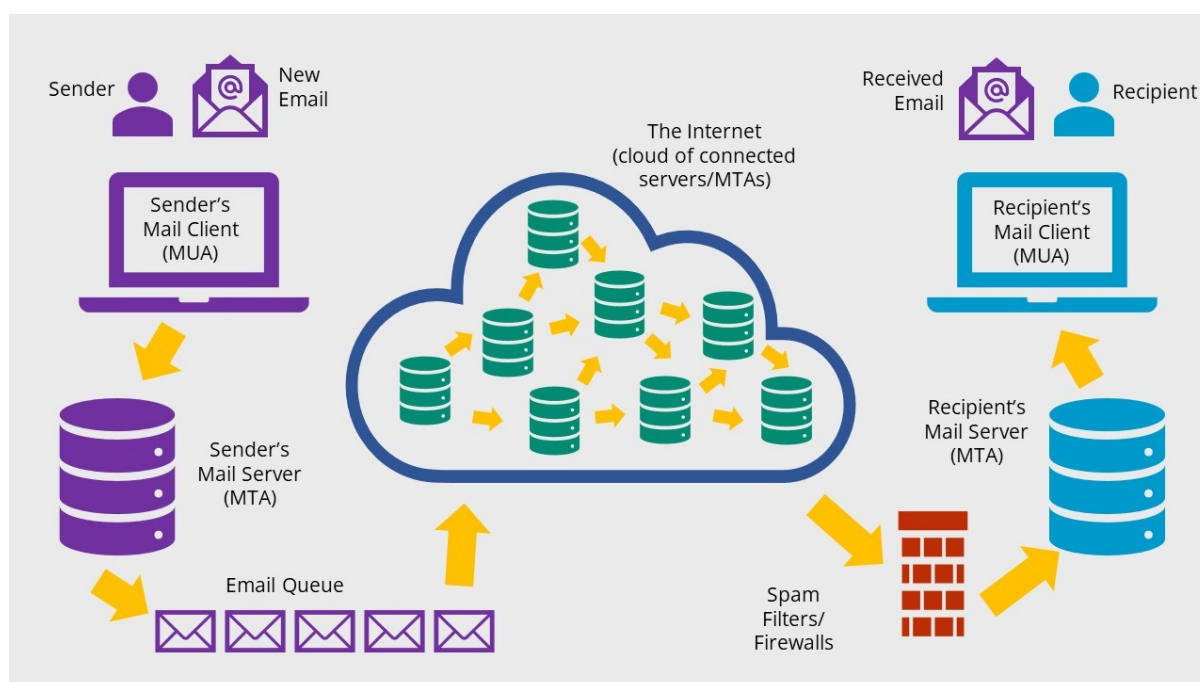


Illustration of how an email is transmitted from the Sender to the Recipient(s)

Module 2.2: What's in An Email

Introduction

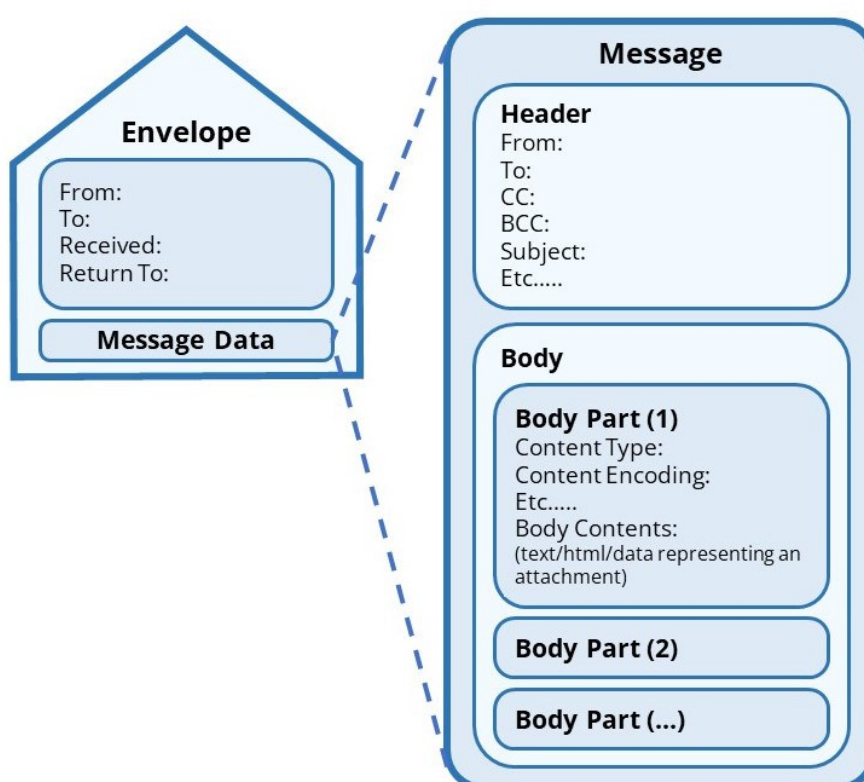
The main elements of an email message are familiar to us all:

- The To, Carbon Copy (CC), and Blind Carbon Copy (BCC) fields for specifying recipients;
- The Subject line, and;
- The main body of the message.

However, when thinking about preserving email, it is important to understand the data model that sits behind what we see onscreen. The data model both provides us with opportunities that will aid preservation, and with challenges we will need to consider.

Main Components of an Email

For any email there are generally up to four main types of components included. The message will be made up of a header section and one or more body sections. These may also be accompanied by attachments. For transmission between the email sender and the recipient(s) these are then packaged into an envelope.



Standardization of Email

There are well-established standards that specify the structure of these components. They ensure the interoperability across systems that has helped make email such a ubiquitous format. Most standards are also extensible which allows for local customization.

Most email standards are managed by the Internet Engineering Task Force (IETF), with RFC 5321 (Simple Mail Transfer Protocol) and RFC 5322 (Internet Message Format) being the two main standards relating to email.

Over the next few sections, we'll look at each of the components of an email in more detail and talk more about the relevant standards.

Envelope

An email envelope is used to “package” the email message for transfer from the sender to the recipient(s) and contains the required information to facilitate that transfer. The structure of an email envelope is defined by RFC 5321, section 2.3.1.

At a minimum, the email envelope must specify an originator (sender’s) address and one or more recipient addresses. These are usually sourced from the email header. Optional extension information can also be added by the system.

As mentioned in the previous module, the envelope is added at the time an email is sent and is normally stripped away before it is delivered to the recipient.

Header

Email headers are also defined within RFC 5321. They consist of a number of header fields, each structured as the field name, a colon, and then the data.

- The fields generally fall into three groups:
- Originator fields: From, Sender, Reply-To, Subject
- Destination Fields: To, CC, BCC

Trace Fields: Message ID, Time and Date Stamps for documenting key stages in the emails transfer, and more.

The Message ID is also defined by a standard (RFC 2392). They are, theoretically, unique and are used to relate emails to a thread they belong to. Although they are defined by a standard, how they are constructed by the messaging application is not. This means their structure can be used to identify which messaging application (e.g. Outlook) generated them.

The metadata contained within the header will likely prove useful when preserving email, providing key information and context. Though it is important to remember that the address in the destination fields might not match the actual recipient(s) of the email. This might be due to the use of an email alias, or a message sent to a mailing list.

Body

As mentioned earlier, an email can include one or more body sections. Body sections are defined by RFC 2045 and will contain the message text and any inline elements, such as images and links.

As with other email elements, while the basic structure of the body is standardized, its implementation is extensible, and so differs between messaging applications. For example, some applications will allow formatting of text using HTML mark-up, while others will only allow plain text to be used.

This often means that there is more than one version of each body section of an email transferred between the sender and the recipient(s). A plain text version is always sent, and one or more formatted versions may also be included.

Attachments

It is possible to attach virtually any type of file to an email, within the size limits imposed by the email clients and systems in use. To facilitate transfer of this wide variety of file formats, attachments are converted to the Multipurpose Internet Mail Extensions (MIME) format and included in the message as body parts.

MIME is an internet standard that allows content encoded in another encoding standard to be expressed as ASCII, the encoding standard used for email. So, for the example of attachments, the binary of the file will be converted to ASCII for transmission and converted back to the original binary file when received.

MIME is also used to convert emails in character sets other than ASCII (e.g. Cyrillic or Arabic) for transfer.

Wrap-Up

In detailing the four main elements of an email, the envelope, the header, the body, and attachments, we have begun to explore some of the standards relevant to the structure and transmission of email.

The next module will introduce a few other standards it is useful to be aware of.

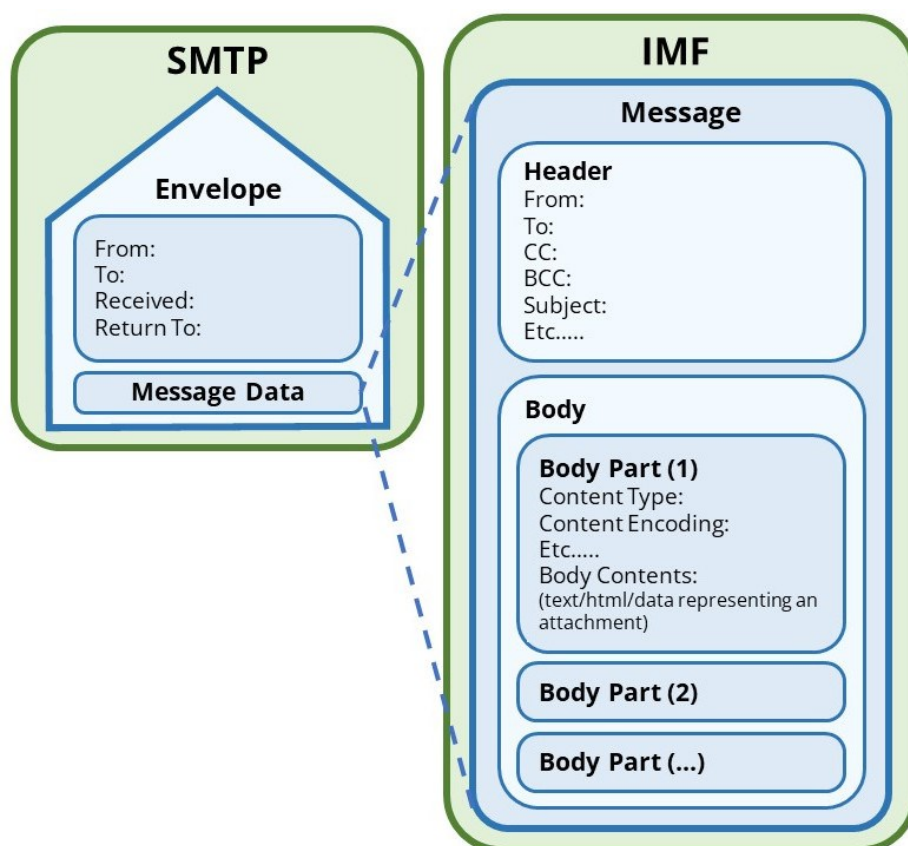
Module 2.3: More on Email Standards

Introduction

We introduced some of the main standards that govern the structure and format of emails in the previous module. In this module we will highlight a few more relating to the structure and transmission of emails.

These standards relate to:

- Message transport - SMTP
- Message structure - IMF
- Message retrieval - POP3 and IMAP



SMTP

The Simple Mail Transfer Protocol (SMTP) defines the method by which email is sent and received. The protocol focuses on the structure and interoperability of the email envelope and is based on the ASCII encoding standard. It does not define any requirements or structures for the content of the email message itself.

SMTP facilitates transfer from the Mail User Agent (MUA), also known as the email client, to the Mail Transfer Agent (MTA), and then between MTAs as the email makes its way to the intended recipient(s).

A different protocol is used for the final delivery of the email, which will be covered shortly.

IMF

The Internet Message Format (IMF) and MIME, introduced in the previous module, together define the content and structure of the email message within the envelope. Like SMTP and MIME, it is based on the ASCII encoding standard.

All emails exchanged through MTAs must use SMTP and IMF. However, emails stored locally on the sender or recipient's system can use a local implementation format. This is often based on the IMF but does not need to be and may use a proprietary format.

Often the locally stored versions include additional metadata that may prove useful for preservation. For example, this might include labeling or tagging, spam scores, or the results of an authentication process.

IMF is also the basis for two of the key preservation formats, MBOX and EML, which we will introduce in a module later in this course.

POP3

The Post Office Protocol (POP3) is one of the two delivery standards for protocols (the other being IMAP) that can be used to transfer an email message from the final MTA to the mailbox of the correct recipient(s). These standards specify how email clients may connect with the mail server to allow the user to view, manage, and delete received messages.

The POP3 protocol allows the email client to connect to the server, retrieve all messages, store them locally, and then delete them from the server. It is largely a one-way process, with the client only able to download and delete from the server, and not update the messages there.

POP3 was developed to serve the needs of users with temporary internet connections, such as dial-up, who would need to be able to work with their email offline. While POP3 is still in use, it has been largely replaced by IMAP which allows more flexibility.

IMAP

The Internet Message Access Protocol (IMAP) was developed to facilitate management of a single email mailbox across multiple clients. This could be a user who accesses their mailbox at different times using a desktop application, a mobile app, and webmail.

IMAP allows an email client to not only access, download, and delete messages, but it can also update the emails it stores. Examples include the functionality to move where it is stored, e.g. to another folder, and the ability to update the status of an email, so if it is read using a mobile app it will later appear as “read” when the mailbox is checked via webmail.

With IMAP, the version of the email saved on the server is considered by the system as the main copy, with any versions downloaded to a local client being a temporary cache. This can be useful for preservation as it means the contents of one or more mailboxes can be downloaded directly from the server rather than having to manage the process through an individual email client. This functionality is included in a number of email preservation tools.

Wrap-Up

Understanding the standards used for the transmission and receipt of emails can help with making preservation decisions. These standards can inform when we decide to capture email, and what formats and processes we will adopt.

In the next module, we will look at another piece in the puzzle: types of email clients.

Module 2.4: Email Clients

Introduction

An email client is the software that is used to read and send email messages by an individual user. It passes messages to and collects messages from an email server (Mail Transfer Agent). As mentioned in the module “How Email is Sent”, email clients are also known as Mail User Agents.

In this module we’ll take a look at the two main types of email client, the types of operations they allow, and how they relate to email preservation.

What is the Main Purpose of an Email Client?

Email clients have two basic roles:

1. To allow users to interact with the email in their mailbox
2. Transferring emails to and receiving them from the email server

As such an email client operates as the intermediary between the user and the technical infrastructure that allows email to function.

Types of Email Client

There are two types of email client in common use: application or app-based clients and webmail.

An application-based email client is one that is installed on a computer, tablet, phone, or other device. It typically downloads a copy of emails to local storage, which allows the user to access them when offline. Common software/app clients include Outlook, Thunderbird, Mailbird, and AppleMail.

Webmail is accessed via a web browser, which acts in combination with a web server as the email client. Common webmail clients include Gmail, Yahoo! Mail, and Outlook.com.

There are some who would argue that using a webmail client is preferable to an app as you are interacting directly with the emails being held on the server. This does, however, generally come at the sacrifice of functionality, as software/app clients normally often have a wider range of options for managing your mailbox.

What Does an Email Client Allow a User to Do?

Email clients allow users to create and send new messages, as well as read, organize, reply to, and forward received messages. Functionality for the organization of emails with a mailbox typically includes options such as folders and/or labels.

Email clients usually also include integrated search engines for finding messages, as well as readers for many common types of attachments (e.g. images or PDFs).

App-based clients like Outlook often offer additional services like a calendar, while webmail services like Gmail allow integration with external calendar apps. Indeed, integrations with other apps and services are becoming increasingly common in email clients.

Email Clients and Preservation

Awareness of how email clients work, and which clients were used by the owner of an email account can influence how you approach preservation. Two examples of issues this might influence are how you will access the account to capture content for preservation and what content you will capture.

If the account holder uses a free webmail service such as Gmail, the only option for capturing content for preservation might be direct access to the account itself. Whereas, if the account holder uses an organizational email account, you may be able to capture content directly from the server using an API and an email preservation tool. The latter will likely require collaboration with IT colleagues.

In relation to capturing content, if the account holder has used a desktop client such as Outlook, and all the available additional functionality, this may provide the opportunity and complications of capturing additional contextual information. For example, will you choose to capture just the email messages, or will you aim to capture calendar data and a custom labeling scheme they have implemented?

Wrap-Up

In this module we have begun to think about email clients and their role in the email preservation process. We will return to the topic of email clients in the future modules on capture processes for select email clients.

Next in this course we will turn our attention to email formats for preservation.

Module 2.5: Email Preservation Formats

Introduction

As with most other areas of digital preservation, there is not a definitive, ideal format for the preservation of email. Rather there are a group of commonly used formats, which each have their own advantages and disadvantages.

It was mentioned in an earlier module, that while emails must use the standard SMTP and IMP formats for transmission, the format used locally by the email client can be custom and/or proprietary. Therefore, you may experience a range of different email formats and need to make decisions about which formats to use for preservation.

In this module we will look at some of the most common formats in use, and mention a few others you may come across.

Format Types - Open or Closed?

As with other types of format, a key issue to consider in relation to formats for email preservation is how open and well documented the file format is. This will likely be a key factor in your choice of format for preservation. To recap from the original Novice to Know-How learning pathway, most formats fall into one of the following categories:

Open Source

Open source file formats are developed through an open community-driven process. The specification is publicly shared intellectual property and is often maintained by a standards organization.

Proprietary but Open

Some file formats are developed by commercial companies, often linked to particular software they develop. With “proprietary but open” formats, the company controls the format and owns all intellectual property but makes the specification available (sometimes with restrictions)

Proprietary and Closed

These file formats are also developed by commercial companies, but the specifications are more closely guarded as a way to control the technology and ensure their market share. With “proprietary but closed” formats, the company does not make the specification accessible and the format can normally only be used with licensed software developed by the same company.

If you are only preserving email from your organization, and the same proprietary email client is used by all staff, you may choose to work with that software’s own format.

Format Types - Granularity

Another main issue to consider is how much content is stored within the email format. Commonly used email formats can encompass anything from a single message through to an entire mailbox.

Therefore, the format used for preservation might be influenced by the selection methodology used. Are individual emails being selected based on their long-term information value? Or are selection decisions being made at an account level? Likewise, will access be provided to individual messages or to all or part of an account in bulk?

We will return to these questions later in the course, in the meantime let’s look at some commonly used email formats.

EML

EML is one of the most commonly used formats for email preservation. It is a text-based format that is closely aligned with the Internet Message Format (IMF). It doesn’t have its own specification, instead relying on the IETF’s RFCs that also define IMF.

EML files typically contain a single email message. Attachments can either be embedded within the file using MIME encoding (as they would be for transmission) or they can be stored separately with a reference included in the EML file.

EML is widely supported and can be exported from and rendered by several email platforms including Outlook, Thunderbird, Apple Mail and Gmail. A number of tools are also available for converting other email formats to EML, although any potential data loss during these processes has not yet been the subject of rigorous analysis.

MBOX

Like EML, MBOX is a text-based format closely aligned with IMF and commonly used for preservation. MBOX files, however, typically contain multiple messages, from a specific folder within a mailbox, to all of the emails within an account. Also, as with EML, attachments can be encoded using MIME within the MBOX file or held separately and referenced.

The ability to encapsulate an entire folder (or more) means that MBOX more easily facilitates the capture of the relationships between messages and threads. At the same time, having everything in the same file means that MBOX files are more susceptible to issues with corruption. If one email is corrupted it might result in the whole MBOX file being unrenderable. The files can also become large and unwieldy if they represent multiple folders or an entire account.

As with EML, MBOX is widely supported for export, rendering, and conversion from other formats. As an example, it is the default export format from Gmail.

PST

PST is an open proprietary email format developed by Microsoft. Like MBOX, a PST file can include multiple emails, usually anything from a single folder to a complete account. Unlike EML and MBOX, the format also allows inclusion of other data created within Outlook software such as contacts, calendar items, and tasks.

Due to the ubiquity of Microsoft and Outlook and the availability of an open specification, despite being a proprietary format, some organizations are accepting and/or using PST as a preservation format. It is also possible to render PST files in a number of other clients.

It is important to note, though, that the export format from Outlook for Macs is different. It is called OLM and does not have an open specification.

Other Formats

While EML, MBOX, and PST are the most commonly used preservation formats for email, there are other options available. You may consider these depending on your technological context and the resources you have available.

For example, Microsoft also offers the MSG format for individual messages. It is to PST roughly what EML is to MBOX. It has not been widely adopted as a preservation format, but it is a format that might be received as part of a deposit of material.

Other formats for email preservation that you may encounter for preservation include XML and PDF. The Smithsonian is using XML for email preservation as part of their DARCmail project and there is currently research underway, led by the University of Illinois, looking at the use of PDF.

Wrap-Up

Like all types of digital content, the list of email formats in existence is extensive. Those included in this module are the most commonly used for email preservation and managing email generally. Therefore, they are the ones you are most likely to encounter.

This is the final module in this course, and before progressing to the next course on “Developing an Approach to Email Preservation”, we will have a knowledge check quiz.

Course 3: Developing an Email Preservation Program

Module 3.1: Understanding How Email is Managed

When establishing a new email preservation program or project, an important first step is understanding the context of how the email to be preserved is used and managed.

Understanding the context will help inform:

- How you advocate for the program and its sustainability;
- Updates needed to existing policies, or new policies that should be drafted;
- Identification of the legal issues that need to be addressed;
- Decisions to be made about selection and retention;
- Development of workflows;
- Selection of tools and systems to be used;
- What preservation actions will be undertaken; and
- How access will be provided.

Therefore, it is worth expending some effort on investigating the technical, legal, organizational, and personal factors that will affect your email preservation work. This process might include some or all of the following activities:

- Gathering and analysis of existing policies and procedures that relate to email use, management, and preservation
- Consultation with IT staff to identify which systems and processes are in place to manage email
- Engaging with relevant stakeholder groups such as email users at your organization, potential external depositors of email collections, and records managers who manage current and semi-current email content.

It is particularly important to engage with stakeholders. Often the reality of email management is different from what might be described in policy and procedure.

The modules in this course will examine the issues mentioned here in more detail. They will offer guidance on what questions you will need answered, what issues will need to be considered, and what decisions will need to be made when establishing your email preservation program, to ensure you have a solid organizational foundation and have captured the information you need to begin developing practical processes and workflows.

Module 3.2: Advocating for Email Preservation

Introduction

Effective advocacy is a key component of any digital preservation work and is of particular importance when establishing an email preservation program.

Advocacy for email preservation requires not only making the case for the importance of the work, but often also requires countering active resistance to the capture of emails for long-term preservation. You may experience strong arguments against preservation due to issues of confidentiality and privacy, in addition to the usual concerns from users about the time the process will take.

A robust and considered approach to advocacy will be important to making progress with your email preservation aims. It is essential to also be aware that advocacy will not be a short-term activity, but likely an ongoing process that will continue once you have passed from developing a program into long-term implementation.

Stakeholder Groups

A first step will involve the identification of key stakeholder groups, so that you will be able develop plans for how to engage each. This should also include key messages you need to convey to each group.

Key stakeholder groups might include (but are not limited to):

- Senior managers/executives, who will approve policy and plans, and provide resources
- Other information managers, particular records managers, who may be involved in managing email at other stages of its lifecycle
- Information technologists, who will also be involved in managing email through systems and services, as well as being potential collaborators on the technical aspects of your work
- Internal email users, who are creating the emails and often do not view them as an organizational record
- External depositors whose collections might include email

Advocacy Methods

Once you have established who your key stakeholder groups are, you will need to consider which advocacy and communication methods will be most effective for each. One commonly used option is the creation of an email preservation task force or working group which can allow inputs from a range of stakeholders and will facilitate planning and decision making.

In relation to the groups mentioned in the previous section, the following methods might also prove effective:

- Senior managers/administrators: Presentations at meetings, elevator pitch, business cases, briefing papers (particularly on risks and opportunities relating to email preservation)
- Other information managers and information technologists: Collaborative meetings, training, joint projects

- Internal email users: Training, information sessions, focus groups, surveys, and guidance
- External depositors: Tailored guidance documents, example deposit agreements, inclusion of email on checklist for deposit negotiations.

We'll discuss engaging with information technologists and the email user stakeholder groups in more detail in a later module.

Key Messages

Establishing a set of key messages about your aims for email preservation will also help facilitate advocacy. They will help to ensure that messaging is clear and consistent, and that the same foundational messages will reach all audiences.

Your key messages will be dependent on your local context and what you are hoping to achieve, but they will likely touch on some of the following:

- The importance of email as a record to support organizational activities, meeting legal requirements, and as a historical document.
- Preservation must be proactive to reduce potential risks, particularly in relation to non-compliance with regulations such as GDPR.
- Effective email preservation can bring opportunities, including increased efficiency, access to improved informational resources, the ability to respond quicker to enquiries such as FOI requests, and potential savings in areas such as digital storage.

Once you have established your core key messages, it will then be possible to tailor these as needed for different stakeholder groups.

NOSE

An important element of successful advocacy is the ability to establish an emotional connection between your audience and the issue you are advocating for. There are several methodologies to help you craft messages that will achieve this aim.

One such method is NOSE, which set out the elements you should include in your messaging:

- Need - empathize with your audience and show them that you understand their own needs
- Opportunity - introduce the opportunities that email preservation can bring about, how it can meet the need in question
- Solution - identify what would be required in order for the solution to be realized
- Evidence - use facts and figures, graphs and charts to show how

Using these elements to craft messages will help ensure impact on the audience, and tailoring key messages to particular stakeholder groups.

Nothing Helps Like a Mandate

One of the biggest aids to advocacy is ensuring there is a clear mandate within the organization for the preservation of email. For most organizations this will involve the development of policy that is officially endorsed by senior management/executives.

This should, in turn, be supported by a clear collecting policy covering deposits of emails from both internal and external sources.

In the next module we will look at policy issues for email preservation and why it is an important element of any program.

Module 3.3: Developing Policy

Introduction

The importance of developing robust and relevant policy for digital preservation is well established. A good policy will set out a clear framework for the development of digital preservation activities, providing a foundation for all decision-making. There are a number of resources available to aid in the development of policies, including the DPC's 'Digital Preservation Policy Toolkit'.

It is also important to remember that some organizations will use different names for policy documents, perhaps calling it a strategy. For the purpose of this module, we are referring to the high-level document that sets out aims and requirements for a program of work.

Developing an email preservation program will likely require creating or updating policies specific to email. Many organizations will have policy relating to email in use, but little to no policy relating to the long-term retention of email. In this module we will examine important issues relating to policy for email preservation.

Part of a Policy Landscape

The first step is understanding where email preservation policy will sit within the current policy landscape of your organization. This may be information you have already collected when researching how email is currently used and managed within your organization.

You may need to examine policies covering issues such as:

- Collecting
- Digital Preservation
- Data Protection
- Confidentiality
- Risk Management
- IT Infrastructure
- IT Security

Through this process you will be able to identify if a stand-alone email preservation policy will be required, or if policy relating to email preservation should be incorporated into existing policies. For example, there might be email retention policies within IT that call for deletion of email after a set period. This would be detrimental to any program aiming to capture email for permanent retention. For collecting from external depositors, a separate email collecting policy, or policy sections, might be useful to address the particularly complex issues relating to rights and compliance.

Development Process

The first step of developing any new policy is advocacy and stakeholder engagement: making sure the correct people are engaged with the process and understand why it is important. Next, you should establish a clear purpose for the required email preservation policy, to ensure the development process stays focused.

The purpose will also help steer the research phase, where you will investigate current good practice. Seeking out examples of email policies for other organizations can be particularly useful at this stage, but it is important to remember that every context is different. Some elements of other organization's policies may be relevant, while others are not.

Drafting of policy will follow, and this may need to be a collaborative and iterative process depending on how policy will be structured. This is particularly relevant if updates are needed to existing policies, where you will need to work with relevant colleagues on any updates. This might be a challenging process as some individuals might be territorial about policy they see as falling under their domain. Being confident in the organization's requirements for email preservation will help, as will gaining support from senior managers.

When developing policy it is important to ensure that the contents are sensible and actionable. Policy can be aspirational in relation to what you are hoping to accomplish with email preservation, but always remember to ensure any goals sets are achievable and realistic.

Once you have draft policy content you will need to have it reviewed by relevant stakeholders and make any necessary updates to develop a version ready for approval.

Approval processes will differ from organization to organization, but generally requires the presentation of the policy documents that have been created or updated for endorsement by senior management. This process is important as this endorsement is often the mandate required to establish email preservation as an organizational priority.

After approval of policy it is important to undertake further communication efforts to ensure widespread understanding of the new requirements as set out in the document(s) and that the requirements are successfully implemented in practice. Depending on the scope of your email preservation activities, the impact may be significant and you may need to utilize change management techniques to ensure adoption.

Policy Elements

When developing policy for email preservation there are several key elements you will likely need to include to ensure it provides the framework you need for practical implementation.

These policy elements will include:

- Selection and appraisal - The types of emails and/or accounts that should be preserved.
- Retention - How long different types of emails/accounts should be retained for.
- Data privacy and security - How these issues will be addressed during preservation.
- Responsibilities - Who will have responsibilities to ensure that the correct emails/accounts are captured and transferred for preservation and when/how often this will happen.
- Access - How will access be provided and will there be an embargo period.
- Preservation - a high-level description of the approach to preservation that will be adopted.

We will address each of the issues mentioned above in future modules and the content of those modules will help inform policy decisions you may need to make.

Using Policy

Once you have approved policy relating to email preservation, it is important to continue the momentum and begin implementation. At a high level this may involve the development of a strategy and/or implementation plan.

From this you will look to develop and test workflows, select tools and systems, document procedures, and create guidance. At all times the approved policy should be the benchmark used for all decision-making.

It is also important to remember that policy should not be a static document. It is important to undertake regular reviews to ensure that it continues to meet the organization's requirements and good practice.

Module 3.4: Promoting Good Practice for Email Management

Introduction

As part of your work to advocate for email preservation and to establish policy, there are two groups to which you will likely need to give special consideration. These are colleagues within your organization's IT department who manage the email infrastructure and services, and email users who are generating the messages you will wish to preserve, both internal to your organization and potential external depositors.

In this module we will take a brief look at each group and the key issues that will need to be addressed.

Working with IT

Your organization's Information Technology department/team will be a key driving force behind how email is currently managed. It is, therefore, essential to ensure that they are engaged with any email preservation program. A good working relationship with IT colleagues can be key to success.

Awareness raising and education on the importance of email preservation will likely be an aspect of relationship building. In particular, many IT teams may already consider themselves to be undertaking email archiving, due to a difference in how the term is used. In IT terms, archiving normally means removal of digital content from live systems to "archival" storage areas that are cheaper to maintain. Clarification of the broad range of activities required to archive email from a preservation perspective will be required.

As mentioned in the introduction to the learning pathway, many organizations also have policies and processes for the auto-deletion of email in order to avoid risks and save storage space. You will need to work with IT colleagues to identify where this might be happening and develop new policies and processes to ensure no emails of long-term value are lost. The need to retain email may put additional pressure on IT resources, so there may need to be negotiation and the development of business cases for additional resourcing.

There will also be elements of developing email preservation workflows that will require support from IT colleagues, such as tool or system procurement and implementation. Again, collaboration on the development of approaches will ease this process. Indeed, a positive working relationship with IT colleagues can ensure more efficient and effective management of email across its lifecycle.

If your organization currently outsources management of the technical aspects of email then direct engagement and collaboration might not be possible. In this case you will need to lobby so that requirements that support email preservation are included in any service contracts.

Working with Internal Users

Engaging email users with the importance of email preservation can also have a significant positive impact on any program. Their input can aid with selection and appraisal decisions, and increased engagement can lead to the capture of more email.

Users often view email as a transitory medium that is disposable rather than an important organizational record. Many also have a strong sense of ownership over the contents of their email account which might clash with preservation objectives.

Guidance supported by information sessions and/or training can help to address these issues. Ownership should be addressed by situating email as an organizational record, owned by the organization, that needs to be managed similarly to other records types.

It is also important to set out clearly what email needs to be preserved, as well as who has responsibilities for ensuring the email is transferred for preservation at the appropriate time. Deletion of email should be strongly discouraged unless following specific guidance for removal of non-record messages (e.g. messages from mailing lists or personal emails).

It is important that policy reflects procedures for good email management and does not encourage unwanted behaviors. User education will also likely be an ongoing process to ensure new members of staff are briefed and that staff continue to adhere to guidance.

Included within the resources of the learning pathway is a guide to email preservation written with record creators in mind. The contents of this resource are available for reuse under a Creative Common license and you are welcome to use the text as needed within your own resources.

Working with External Depositors

When working with email users outside of your organization the potential to influence their email management practices is somewhat reduced, but it is still possible to offer guidance on good practice. If you are expecting to receive future deposits that will include email, it is worthwhile raising good email management practices as you build relationships with potential depositors.

It will likely also be useful to develop a version of email management guidance aimed specifically at those using email for personal purposes, offering guidance on what types of emails they may want to retain and what can safely be deleted. This can then be shared with potential depositors as well as published on your website.

Those using their email accounts for personal purposes will likely be generating more non-record email than email with long-term value, so it is essential to offer clear guidance on what is worthy of retention. This might include encouraging users to file selected emails in a specified folder or folders if it is deemed important.

Module 3.5: Legal Contexts for Email Preservation

Introduction

Identifying relevant legal issues can be a powerful tool in advocating for preserving emails. This offers an opportunity to frame email preservation in a way that will likely resonate with those in senior positions, allowing you to make the case for policy, procedural, and resource requirements.

Legal considerations do, however, also introduce a number of risks that put the survival of email in danger. While information managers will view the retention of email to meet legal requirements as a positive move, others may view the same process as opening up an organization to legal risks.

It is this state of mind that often pushes organizations towards email management policies that favor deletion, whether it be manual or automatic after a fixed period of time. Finding a balance between these two opposing views is essential when addressing legal issues for email preservation.

But what legal issues might you face? While laws will differ from country to country, and as such specific guidance is outside the scope of this learning pathway, there are common areas of law to be considered.

Key Areas of Law

In his Technology Watch Report 'Preserving Email', Chris Prom identifies four key areas of law that might need to be considered in relation to email preservation:

1. Public records and freedom of information (FOI) laws
2. Industry-specific laws and regulations
3. Rules relating to discovery/disclosure in legal cases
4. Data protection and privacy laws

To this list we will also add issues relating to intellectual property and moral rights.

The relevance of these areas of law will depend on your own organizational context, but it is worth considering each and determining if and how they will impact your email preservation work. Each has the potential to influence decisions made about, and processes established, for selection, appraisal, retention, and access.

Over the next few sections, we will introduce each of these areas and their potential impact on email preservation.

Public Records and Freedom of Information

Where public records legislation exists, most laws either clearly establish or imply that email is a public record. For relevant organizations it is, therefore, a requirement that it is managed as such. This, of course, will also require organizations to provide access to pertinent emails for FOI requests.

The need to meet these requirements can be an important driving force in establishing the need for email preservation, at least in the short-term. Identifying relevant public records and FOI legislation should be part of your planning and advocacy activities.

Industry-Specific

Most industries will be governed by specific regulations relating to record keeping, compliance periods and rules regarding data use. These should be considered when planning your approach to email preservation, particularly when establishing retention periods.

For example, in the financial sector US law requires that audit records and any related correspondence are retained for a minimum of seven years. In the UK the term is six years from the end of the relevant accounting period, with some caveats for records that should be kept longer. These include records relating to assets with a lifespan exceeding six years, in which case records must be retained for the lifespan of the asset.

Legal Cases

Many legal cases now include digital content amongst the evidence provided. Often known in law as Electronically Stored Information (ESI), emails will be covered by discovery/disclosure rules where they exist. These provisions require parties involved in the suit to provide copies of all evidence that is relevant to the claims and defenses of the case prior to appearance in court.

This process may also require proof that the records are authentic, establishing the need for proactive management and preservation of email for its potential evidential value. In some jurisdictions, such as within the US, severe sanctions can be applied if ESI with proven integrity cannot be provided.

These requirements have led to the development of a new branch of law practice known as e-discovery. If the legal evidential value of email is important to your organization, you may wish to seek the advice of an e-discovery expert to ensure processes and documents meet the relevant requirements.

Data Protection and Privacy

While the areas of law mentioned so far offer inducements to retain email, such as avoiding fines, the opposite is true of data protection and privacy legislation. Organizations may, instead, face financial and reputational risks if personal or sensitive data is shared without the proper permissions and authority. This can often lead organizations towards an overzealous deletion policy.

This is further compounded by the fact that most data protection and privacy legislation, while drafted with the processing of digital data in mind, does not make accommodations for how that data is structured, shared, and managed in formats such as email. This is true of the European General Data Protection Regulation (GDPR), which includes clauses such as “the right to be forgotten” and offers little leeway for issues such as handling emails from third parties.

The ramifications of this legislation are not yet fully known as it has yet to be fully tested in the courts. In the meantime, selection and retention decisions will need to be made to balance restrictions with the potential long-term value of email records. These issues will likely be exponentially more difficult when working with emails from external sources, where less might be known about the contents of an email inbox and those who have sent emails to the

account. Establishing an approach to managing emails that addresses the issue of privacy and takes account of relevant legislation is therefore essential.

Intellectual Property and Moral Rights

The final area of law to be considered relates to Intellectual Property and Moral Rights. Consideration of these rights introduces some of the most complex issues in relation to the preservation of email, not least because some issues would encourage the retention of messages, while others their deletion.

Email sent and received within the same organization will likely cause few problems in terms of capture as most employment contracts include language that establishes IPR for any work products as belonging to the employer, which will include email. Issues may, however, arise in relation to providing access to email collections that include significant intellectual property, such as designs included as attachments. For this reason, some organizations will institute a standard embargo period on access to emails.

When considering emails from senders outside of the organization, more issues might arise with regards to IPR. For example, content such as designs may have been shared in emails but the other organization specified they could only be shared with certain individuals. This might recreate the need for tailored selection, appraisal, and retention approaches.

These issues are likely only amplified when considering email from external depositors. Within the more informal setting of personal email, correspondents are more likely to share content where the IPR and moral rights are held by a third party. For example, photos taken by a friend at a social event, or a song or video they wish to recommend. Also, the senders of emails received will likely have no knowledge of their messages being shared for preservation. This introduces complexities in relation to their moral rights in relation to reuse and sharing of their intellectual property. For this reason, extra care will need to be taken in addressing IPR and moral rights issues with deposits of personal email.

Wrap-Up

Each of these legal areas has the potential to impact on your email preservation work to a greater or lesser extent depending on your organizational context.

As mentioned earlier in this module, they have the potential to affect activities such as selection, appraisal, retention, and access. In the next module we will focus on one of these in particular: retention.

Module 3.6: Retention and Deletion

Introduction

Issues relating to retention and deletion have already been mentioned in earlier modules. Decisions on what you will aim to keep and what can be deleted will be foundational to your approach to email preservation.

As has already been identified, it is also an area where proactive efforts will be required to prevent the loss of email to auto-deletion policies and processes and overzealous pruning of individual email accounts by users.

A first step will likely be addressing how the word retention is used by different email stakeholders. In records management and archiving, “retention” is used to describe the **minimum** amount of time something will be maintained, whereas the opposite is often true in technology circles, where it will determine the **maximum** time content will be retained.

In this module we will consider some of the key retention issues you will face, different approaches to deploying retention plans, defining responsibilities, and the importance of testing your retention plans and processes.

Key Issues to Consider

When considering retention of email there are four key issues to be addressed:

- How does email relate to existing retention schedules and processes? Will a new schedule be required or can email be accommodated with updates to existing schedules?
- How long will email be retained? It is likely that there is no singular answer to this question. Different types of emails will need to be retained for different periods of time.
- When will email be captured for retention? At the point of creation? Receipt? At regular intervals? When an employee leaves their position?
- Who will have responsibilities for carrying out retention plans? Will individual users have the power to make decisions? Will responsibilities lie with records managers and archivists? Or is there any scope for some automation of the process?

Answering these questions will be essential in developing your email retention plans and related schedules and guidance.

Retention Periods

There is often no single answer to how long email messages should be retained. Exact plans will be dependent on your organizational context, but emails do tend to fall into three broad categories for retention:

1. Emails of a transactional nature that can be routinely deleted. These are often emails relating to the day-to-day operations and administration of the organization's work. This category may also include spam/junk mail or messages such as those received from mailing lists or of a personal nature.
2. Emails with short to medium-term value to meet regulatory and legal requirements. These are emails that are required as evidence of particular actions or programs of work as set out by relevant regulations or legislation, usually with clear retention periods. As discussed in the module on legal issues, this might include messages such as those produced as part of a financial audit. Another example might be within the manufacturing industry for messages relating to an individual product that need to be retained for its lifespan.
3. Emails with long-term historical value. These are likely to be the smallest group of emails to be considered. These will be messages that document key decisions, projects, or activities carried out by the organization or individual. Within an organization these are often more likely to be created or received by senior members of staff.

Responsibilities for Retention

In addition to setting out retention periods, it is also essential to identify who will have responsibilities for implementing those plans for retention. With email preservation, those responsibilities might sit with one or more of the following groups:

- Email account users
- Records managers
- Archivists
- Information technologists

Alternatively, they may be automated to some extent depending on the sophistication of the technological solutions deployed and the complexity of the retention decisions and periods.

If deletion by email account users is to be allowed, clear guidance, training, and support will be crucial.

Decisions on how and when to capture emails are also relevant and we will return to that topic in a future module.

Developing and Testing Retention Plans

When developing retention plans you will also need to decide what information will be used to help implement the requirements they outline, i.e. how to identify which emails fall into which category for retention. Information that might be used can include:

- The name/position of the account holder
- Email exchanges between certain correspondents
- Particular folders within an account (e.g. only "Sent" mail)

- Email subject lines or content.

Each option brings its own advantages and disadvantages.

Whatever approach is chosen, it will also be worthwhile testing out implementation of the retention decisions, perhaps as a pilot project before rollout across a whole organization. It should also be monitored over time using robust quality assurance practices.

Module 3.7: Thinking About Email Preservation Workflows

Introduction

As is the case when preserving any type of digital content, it is good practice to invest time and effort into developing and documenting clear workflows for email preservation. This will help create efficiencies, establish a baseline of good practice, and ensure consistency, accountability, and sustainability (i.e. processes do not just “live” in one person’s head).

In upcoming modules, we will look at key elements of email preservation workflows and how they might be implemented. But before we delve into specifics, we will look at what workflows you may need to develop and how they might be documented.

What Workflows Are Required?

Your approach to developing workflows for email preservation will depend on a number of factors including:

- Your key aims and objectives for email preservation, e.g. do you plan to establish basic workflows now that you will develop further over time?
- Relevant policies, e.g. are there particular requirements set out within the policy?
- Resources you have available, e.g. will you be developing workflows that can be implemented by an individual member of staff, or will the work be collaborative? Will you have funding to invest in new infrastructure to support the work?
- The technology you will be able to use, e.g. do you already have a repository system in place that supports email preservation? Does your organization support the deployment of open-source tools? Is there sufficient storage space available?
- The types of email content you will be working with, e.g. will you just be preserving emails from within your own organization and/or will you be preserving emails from external depositors?

The final point above will have a significant impact on the steps involved in any workflow, particularly in relation to the selection, acquisition, and capture of emails. Additional steps relating to negotiation and the preparation and signing of agreements represent some of the additional considerations when developing workflows for preserving email from external depositors. If you will be receiving emails from multiple sources you will likely need to have different versions of workflows tailored to the specific needs of each situation.

You will also need to consider if you will develop workflows across the whole email preservation lifecycle, or if you will create separate workflows for different elements. In this learning pathway we will be breaking down the lifecycle into the following elements:

1. Selection and Capture
2. Appraisal and Processing
3. Preservation
4. Discovery and Access

You may wish to develop workflows for each of these, or a single workflow that documents activities across all four. As with most things within digital preservation, the option you select will depend on your own organizational context.

The “Future of Email Archives” report mentioned earlier in the learning pathway includes a number of example email preservation workflows you may wish to examine before developing your own. The report is careful to add the caveat that there is no perfect, one-size-fits-all email preservation workflow. Each organization must develop the right workflow for their context.

Documenting Workflows

There are many different approaches to documenting workflows, and you may find there are already established practices within your organization that you can follow. If not, we would encourage you to consider using one or more of the following:

- Workflow diagrams
- Written procedures/guidance
- Workflow checklists

Workflow diagrams provide a high-level graphical representation of the steps included in a workflow. These are initially useful during the workflow design process to help with visualizing how elements of your email preservation activities will fit together. They can help facilitate a structured approach to identifying requirements, selecting tools, and assigning responsibilities. They can also be a useful tool when explaining the importance of email preservation and how it works, providing a foundation for advocacy and collaboration.

Written procedures and/or guidance can describe the step-by-step implementation of the workflow represented within a workflow diagram. Procedures are important to ensure consistency when different staff members are implementing the same workflow. They also ensure sustainability of the program, so that the knowledge does not reside with one individual. They help establish accountability, showing that the organization’s approach to email preservation follows good practice. Procedural documents may be required if the organization plans to apply for any form of accreditation or certification.

Finally, to accompany procedural documents, it can also be useful to maintain checklists for the steps carried out within a particular workflow. These, again, help to ensure that workflows are carried out consistently, acting as a memory aid for those undertaking email preservation

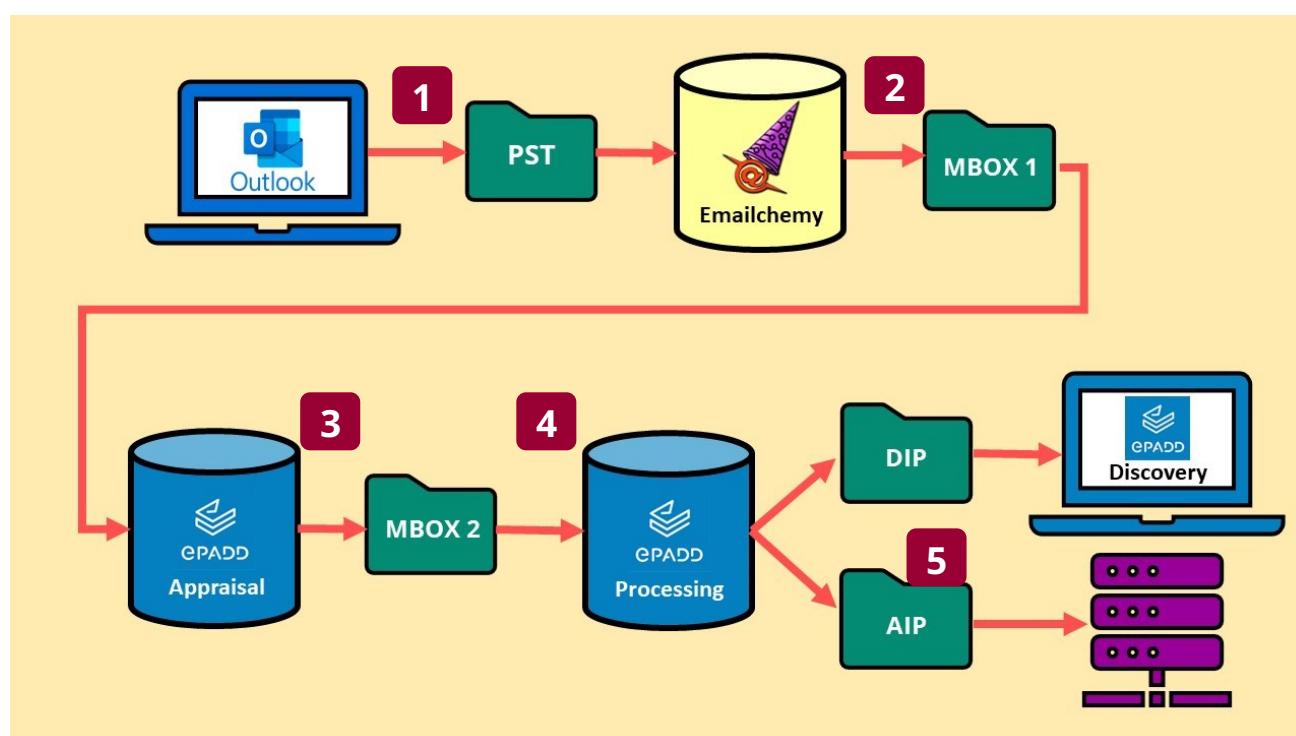
activities. Checklists should break each step in the workflow down into its most granular level so those undertaking the work can check-off each activity when it has been completed.

If you do decide to document your workflows, please do consider sharing them with the wider community if possible. The Community Owned Workflows section of the COPTR tool registry provides a facility for sharing, as well as examples of how others have documented their workflows.

Example Workflow Diagram

The image below provides an example of how you might start to document a simple workflow. It depicts the following steps:

1. The account holder's mailbox is exported from Outlook as a PST file
2. Emailchemy is used to convert the PST file into an MBOX file
3. The MBOX file is imported into ePADD, the messages contained are appraised, and a new version of the MBOX file is exported.
4. The new MBOX file is imported into the ePADD Processing module, the emails are processed for preservation and access, including the addition of metadata.
5. After processing two versions of the email archive are exported: a Dissemination Information Package which is imported into the ePADD Discovery module to help facilitate access, and an Archival Information Package which is transferred to archival storage.



Module 3.8: Introduction to Email Preservation Tools and Systems

Introduction

There are a wide variety of different tools and systems that can be used as part of your email preservation workflows. The tools and/or systems you decide to deploy will depend on a combination of factors including your policy, needs, technological capabilities, and the resources you have available.

In this module, we will introduce the range of tools and systems available to facilitate email preservation activities.

Types of Tools and Systems

The tools and systems available to aid with email preservation generally fall into one of the following categories:

- Commercial email management tools that include functionality that can contribute to email preservation
- Digital forensics tools that can help with processing emails for preservation
- Tools designed specifically for email preservation that address one or more elements of the life cycle
- Repository systems that include functionality to manage email across the life cycle.

Each of the tools and systems have their strengths and weaknesses, and you may find that you need to use a combination of them across the email preservation life cycle to meet your needs.

A list of the tools and systems mentioned is included with the downloadable resources that accompany the learning pathway. The list provides more details on each tool, including where in the life cycle they are most commonly deployed and the technical requirements for implementation. This list is not exhaustive and information on more tools can be found on the website for the [Task Force on Technical Approaches for Email Archives](#) or in the [COPTR Tools Registry](#).

Commercial Tools

There are a number of commercial email management tools that you may find useful as part of the email preservation lifecycle. Some of the most commonly used are Aid4Mail, Emailchemy, and MailStore.

These tools have been developed to support email “archiving” activities commonly carried out by information technology professionals. The types of tasks they facilitate include capture of email from live systems, conversion between different formats, deduplication of messages, metadata extraction, and transfer to “archive” storage.

These tools can be especially useful during the acquisition and processing phases of email archiving. In particular, they are most often used to facilitate “normalization” of emails from proprietary formats to selected preservation formats, a process that is not well-supported by free tools.

The major drawback of using these tools is, of course, the cost required to obtain a license for use. The price of commercial tools varies significantly and, if you need to use one of these tools, your choice of which one to deploy may be dependent on the funding you have available and the number and type of license(s) you will require.

Forensic Tools

Digital forensics tools were originally developed to support the work of law enforcement bodies and to provide evidence for legal cases. They focus on the recovery, investigation, examination, and analysis of material found in digital devices. Their analytic functionality has, however, found a secondary use in the digital preservation community.

In particular, they are often deployed to allow the creation of disk images, identification of “deleted” files, format conversion, malware detection, characterization, metadata extraction, and the creation of fixity information. The two most commonly used tools amongst those working in digital preservation are BitCurator, an open-source system developed with digital preservation in mind, and Forensic Toolkit (FTK), a commercial product developed for forensic investigation.

Digital forensic tools are most often deployed with the email preservation life cycle to aid with acquisition, appraisal, and processing. In addition to the uses mentioned above, they also support the extraction of attachments from emails.

Email Preservation Tools

A number of tools have also been developed specifically for the purpose of email preservation. These include DArCMail, ePADD, and RATOM. Each has different functionality shaped by the reasons for their development.

RATOM (Review, Appraisal, and Triage Of Mail) primarily focuses on the appraisal and processing of email for preservation. It supports the implementation of quite advanced and powerful workflows that can include automation and NLP. It does, however, need reasonably advanced tech skills to utilize including database implementation and python scripting.

DArCMail has been developed to support elements of appraisal, preservation, and discovery and access. It utilizes an XML-based format for preservation and delivery of email content and supports normalization to this format, as well as support for tasks such as bulk processing of up to 100,000 emails in a batch, and ongoing integrity checking. Like RATOM, DArCMail requires the implementation of a database and python scripting to deploy.

ePADD is a browser-based tool that supports the appraisal, processing, preservation, discovery, and delivery of emails. Functionality includes exporting attachments, NLP and NER support for appraisal and processing activities, and options for redaction. Deploying ePADD requires only a suitable web browser and the installation of Java.

Due to the lower technology requirements for the use of ePADD, it has been chosen for the tool demos included for this course. This is not, however, a recommendation of the tool above the others mentioned in this module. It is always important to consider which tool is most appropriate to meet your needs. All three of the tools mentioned in this section are open-source, are available for free, and are accompanied by guidance on implementation and use.

Repository Systems

Most digital repository systems will also include support for preservation of email within their functionality. This is often provided by a microservices approach where individual tools are combined to offer a complete workflow managed by the repository.

Access to and use of a repository's pre-designed and tested workflows can bring efficiencies for practitioners, whilst also removing some of the barriers to tool deployment such as the need for skills like scripting. In particular, efficiencies are often gained from the ability to use one system across the life cycle of email preservation, especially as that system will integrate with storage and offer automation of routine actions such as integrity checking.

The drawbacks of repository systems include the financial resources and time required to procure a system, as well as the technical support required for implementation of non-cloud-based options.

Repository systems that incorporate email preservation workflows include:

- Archivematica
- Arkivum
- LIBNOVA
- Preservica

A deep dive into each of these repository systems is outside the scope of this learning pathway, but a short description of the functionality each offers for email preservation is included in the list of tools and systems within the additional resources.

Course 4: Selection and Capture

Module 4.1: Selection Methods

Introduction

One of the more important questions you will need to answer in developing your email preservation program is "What do we want to keep?" Considering the sheer volume of email

that is created, the decisions made in relation to what email to select for preservation might have a significant impact on ensuring an email preservation program is a success.

There are a range of approaches that can be taken to the selection of email for preservation, and in this module we will highlight some you may consider. Each comes with its own pros and cons that you will need to assess against your aims for email preservation.

You may find that a single method will be suitable for your context, or you may wish to employ different methods for different types of collections or content. As an example, you might choose one method for email generated by employees of the organization and a different method for collections of emails received from external depositors.

When deciding which method(s) to employ it is useful to consider the following issues:

- What collecting policy and retention schedules already exist that might be relevant to email? Will updates be required?
- How much scope, in terms of time, resources, and relationship building, is there for working with email account holders as part of the process?
- When do you expect appraisal to occur and who will carry out those responsibilities? Is pre-appraisal by the account holder a possibility, and/or will appraisal be carried out by archivists/information managers?

In the following sections we will examine four commonly used methods for selecting emails for preservation.

Keep Everything

The approach that seems most simple on the surface is to just keep everything. This significantly reduces any burden on the account holder and reduces the amount of guidance needed and time spent on support for selection and pre-appraisal processes. For this reason, you may wish to consider the “keep everything” approach when working with external depositors.

However, there are arguments for not keeping everything. The first downside of this approach is that it may exponentially increase the amount of storage required to hold the emails collected. While storage costs continue to decrease, the implication of potentially holding multiple copies of a large number of email accounts can be an obstacle.

The second downside is that this might be delaying the burden of work, requiring significant effort from archival staff later in the acquisition process, or when providing access. Capturing everything opens the door to chances of capturing sensitive information that will introduce risks such as those relating to data protection and confidentiality. These would need to be addressed at some point, either manually or perhaps with the aid of advanced techniques such as natural language processing or machine learning.

Ultimately, this method is a balancing act between the costs of storage, staff time, and potential data sensitivity risks.

Focus on Sent Items

One popular method focuses on the selection of emails contained within an account's "Sent" folder. A key rationale behind this approach is that it captures all of the email generated by the email account holder themselves.

Benefits of this method include a vast reduction in the amount of spam and irrelevant messages captured. Also, due to the fact that most email correspondents default to retaining previous messages from the thread within replies, this will also capture the text from a large number of the messages that were received as well.

Disadvantages include the possibility that important messages received by the account holder will be missed if no response was sent. Additionally, email clients may strip attachments from messages when replying to or forwarding emails, which can also result in the loss of important records.

The success of this approach often relies on an understanding of an individual account holder's email management behaviors, so you can assess if the contents of their "Sent" folder will act as a good surrogate for the email held within their account.

Self-Curation

With a self-curation approach, you would be placing (at least some) of the selection and appraisal decision-making in the hands of the account holder. Self-curation allows the account holder to identify which email messages should be selected for preservation. This might focus on an entire account, particular folders, and/or specific messages.

A successful self-curation approach usually requires a significant investment of effort in building relationships with account holders and providing them with clear, actionable guidance. This can, however, also be an important advocacy activity, particularly when working with those who are nervous about handing over the contents of their mailbox. Allowing them some control over the process will help engender trust.

Issues can, however, arise as a promise to carry out self-curation activities does not always translate into action. You may expend energy working with the account holder, only to find they have not carried out the discussed selection and deletion plans. There can also be problems due to different understandings of what might be useful to those accessing the email in the future. Those outside of information management often have a narrower view of what might be considered a record.

Key Accounts

An increasingly popular method of selection involves the identification of significant account holders and selecting their email for preservation. Those account holders might be senior executives or managers within an organization with decision-making powers, key personnel working on important projects, or potential depositors who have made a notable impact on society.

This method can make selection straightforward, as long as clear criteria for account selection are established. For example, it often significantly reduces the need for appraisal of individual messages.

The disadvantages include the possibility that important emails sent or received by more junior members of staff are missed. Also, depending on the account holder's email habits, a lot of unnecessary email might be retained.

A particular implementation of this approach, known as the Capstone Method, has been developed by the National Archives and Records Administration (NARA) in the USA. We will look at this in more detail in the next module.

Module 4.2: Focus on the Capstone Approach

Introduction

As mentioned in the previous module, the National Archives and Records Administration (NARA) in the USA has developed a specific implementation of the selection of key accounts called the Capstone Approach. The method has proven popular across the digital preservation community, with many other organizations adopting its structure and adapting it to their own context.

In this module we will describe why the Capstone Approach was developed and how it has been structured and implemented. Later in the learning pathway you will also find a case study on how it has been adapted for use by a UK Governmental Agency.

Development of Capstone

In the mid-2000s, NARA identified that while email had long been in regular use across the Federal Government, they had not received significant deposits of related content for preservation. This led to an active program of work that included research, congressional hearings, and ultimately contributed to the "Presidential Memorandum on Managing Government Records", signed by President Obama on 28th November 2011.

The section of the Memorandum covering email required that by "December 31, 2016, federal agencies must manage email records in an electronic format". With agencies needing to deal with an overwhelming volume of email records, a practical and realistic approach was needed. An email working group led by NARA developed the Capstone Approach, which was first

published as a bulletin titled “Guidance on a New Approach to Managing Email Records” in August 2013.

Aims of Capstone

The main aims (and benefits) of the Capstone Approach were as follows:

- Improve management of email across the Federal Government in a way that would facilitate more email records being passed to NARA
- Provide a practical and realistic framework that federal agencies could use to help them effectively manage email
- Offer the guidance and support that would allow individual departments and agencies to select which email accounts should be preserved
- Facilitate the deletion of non-record email after appropriate retention periods
- Reduce reliance on a “print and file” approach to managing email records

There is no mandate that agencies and departments must implement a Capstone approach, they may develop their own suitable alternative.

Capstone Officials

The core element of the Capstone Approach is the identification of email accounts for preservation based on the role or position of the account holder within the relevant agency or department. Therefore, one of the biggest challenges of developing the approach was creating a definition of a “Capstone Official”

In the end, this led to the creation of nine categories of Capstone Officials. In summary, they are:

1. The Head of an Agency
2. The Principal Assistants to the Head of an Agency
3. The Deputies of those in category 1 and 2
4. Staff Assistants of those in category 1 and 2
5. Those in principal management positions, e.g. Chief Operating Officers, Chief Technology Officers, Chief Financial Officers, etc.
6. Directors of significant program offices
7. Principal Regional Officials
8. Those in roles that offer advice and oversight to an agency
9. Those in roles not covered by the above that require a Presidential appointment and Senate confirmation

The guidance also states that anyone fulfilling a role above in an “acting” capacity for 60 days or more should have their email account preserved, and that there is scope for inclusion of those in other roles if their work is related to mission critical functions, policy decisions, or holds historical significance.

Retention Periods

The retention periods for emails produced by those working in federal agencies are grouped into three “Items” within the Capstone guidance.

Item 010 includes the definition of Capstone Officials and states that their email accounts are for permanent retention. There are, however, allowances for the deletion of personal emails and emails such as spam and those from mailing lists.

Item 011 sets a minimum retention period of 7 years for the emails of all account holders not included in either item 010 or 012, although longer retention periods can be applied depending on business needs. The majority of email managed within an agency would be covered by this item.

Item 012 sets a minimum retention period of 3 years for employees who are in administrative or support roles, e.g. those performing very specific administrative or routine duties.

Guidance, Verification, and Approval

NARA provides a range of guidance and support for agencies looking to use the Capstone Approach for managing email. This began with the information contained within the original bulletin, and has been supported by a General Records Schedule, information session, and guidance available on their website.

Agencies are also supported through a verification and approval process for their implementations. They can seek approval of their list of identified Capstone Officials and their plans for implementing Capstone and related records schedules.

Module 4.3: Capturing Email Archives

Introduction

Once you have selected email for preservation, the next questions to answer are when and how you will capture that email? The main goal of this process is to remove the email from where it is being managed by the account holder(s) and system admin(s) and into the custody of those responsible for its preservation.

In this module, we will introduce the key issues to be considered when planning for the capture of emails for preservation and the most common methods used for capture processes.

Key Issues

When planning for the capture of email for preservation there are five key issues to be considered:

- The mechanisms to be used for capture and transfer of the emails
- How security issues will be managed

- How the integrity and authenticity of the emails will be maintained
- When in the email lifecycle messages should be captured
- If the capture method will export a faithful copy of the emails

We will now examine each of these issues in turn.

Transfer Mechanisms

The mechanisms used to transfer email will impact how and when email can be captured, what formats it can be captured in, and if there are any possibilities for appraisal at the time of capture. Options for transfer include capture from other intermediary systems, direct export from mail servers, export from webmail services, export from email clients, and receipt as part of a disk image. Each option offers opportunities and has limitations. You may find you need to use a range of different transfer mechanisms depending on the emails you are capturing for preservation. Later in this module we will look at common transfer mechanisms in more detail.

Security Issues

There are two main issues to consider with regards to security. The first is ensuring the security of the email content during the transfer process. This is particularly important if the messages contain sensitive information and/or if they are being transported using removable media. It is, therefore, important to use encryption where possible, such as purchasing an encrypted hard drive to be used in the transfer process.

The second security concern relates to the higher than normal likelihood that email may include malicious content such as viruses. As mentioned earlier in the learning pathway, email is particularly vulnerable, in comparison to other types of digital content, due to the ways in which it is transmitted from the sender to the recipient(s). This is particularly due to the prevalence of spam messages. To counter this, you may consider quarantining any newly received email content for a set period, as well as potentially carrying out initial processing on a non-networked computer.

Integrity and Authenticity

To maintain the evidentiary value and build trust in email preservation processes you must actively manage authenticity and integrity of the email. As with other types of digital content, integrity checking using fixity tools will be a key mechanism for achieving this. It is important to carry out integrity checks before and after any time the emails are moved and to record information about this process as part of preservation metadata.

When to Capture Email

The options for when to capture email fall roughly into two camps: a rolling approach or capture at the end of use of the email account. A rolling approach might see email captured at regular intervals, such as six-monthly or yearly or it might be a continual process where emails are captured at either the time of sending or receipt. At the other extreme, you may choose to capture email when the account falls out of use, for example, when the account holder leaves

their role with an organization or as part of a donation from an individual of significance who has passed away.

Export Fidelity

Whichever export method is chosen, it is important to ensure that the copy of the email content is a good representation of the original messages. To ensure that content is exported as you'd expect, it is wise to conduct some formal testing and establish quality assurance procedures. An example issue to check for would be if the email is likely to contain messages in multiple languages, you should check that any special characters are not corrupted in the export process.

We will now look at the five most common methods for capturing email.

Export from Intermediary Systems

The first method to mention is the capture of email from intermediary systems such as email "archiving" or journaling systems, or an EDRMS. There are many different options available for "archiving" emails at the time of sending, or for capturing emails as they are received. These may be offered as additional functionality with an email client or as standalone software, integrated with email systems. Journaling systems, in particular, are viewed positively as they allow the capture of email according to predefined rules at the time of its receipt. They also can facilitate the retention of the email "envelope" which is usually discarded upon receipt. With EDRMS, you will be accessing emails that users have specifically identified as records, which may be accompanied by additional metadata.

The benefit of this option is that there will likely have been, at least some, pre-appraisal done. The drawback is that you will likely have little or no control over the format in which email is received, relying on the options available from the intermediary systems.

Direct Export

The next capture method to consider is direct export from the mail server. This process involves using an IMAP connection, similar to the one used by an email client to retrieve email. This is one of the most technically complex options and will likely involve working with colleagues from IT to set up the process and ensure you are able to access content in the format you need. For many, capturing email directly from the server is the preferred option as it does not require working with individual account holders and offers possibilities for some automation and/or high-level appraisal as part of the process.

Web Service Export

Most email web services will offer options for exporting data from email accounts. This is usually a relatively straightforward process and typically allows the selection of either a complete mailbox or specific folders. Drawbacks include the need to have direct access to the

account in question and the likelihood that there will be limited options in terms of the format of exports. The process also usually involves format conversion as part of the process and there has been little research to date on potential data losses that may occur. Potential loss of header information has been particularly identified as a concern. Later in this module we will demo the process for capturing email from the Gmail web service.

Email Client Export

Similar to export from web services, most email client software also allows export of all or part of an email account. The process is usually quite simple and some clients offer a wide range of filtering options that can facilitate significant pre-appraisal. In Microsoft Outlook this includes the option to select individual folders, as well as basing selection on a combination of keyword searching, email addresses, filtering by status and the ability to build specific criteria for a wide range of fields. Export from email clients also often includes the option to capture additional content such as calendar entries, contact information, and task lists. The drawbacks are similar to those of export from a web service, firstly you will need direct access to the email account, perhaps even on the same computer as used by the account holder. Secondly, the formats available for export are likely limited. For example, Microsoft Outlook only allows export in a Comma Separated Values (.CSV) file or in the proprietary Microsoft .PST mailbox format. We will look at the process for exporting from Outlook in a future module.

Disk Imaging

Last amongst the commonly deployed capture methods is the inclusion of email within a disk image taken of the account holder's computer. This method is most often used when capturing the digital content belonging to an external depositor as it allows everything to be captured in a single process. This may be considered to be one of the least desirable options as it removes any options for format selection or pre-transfer appraisal. You will likely also only capture emails that have been stored locally on the computer, and not messages that are only held on the relevant server. It may, however, be the only option available due to time and resource constraints. If email is captured through disk imaging, you will need to think carefully about how to extract the email and if there might be any issues of data loss.

Module 4.4: Capturing Email from Outlook

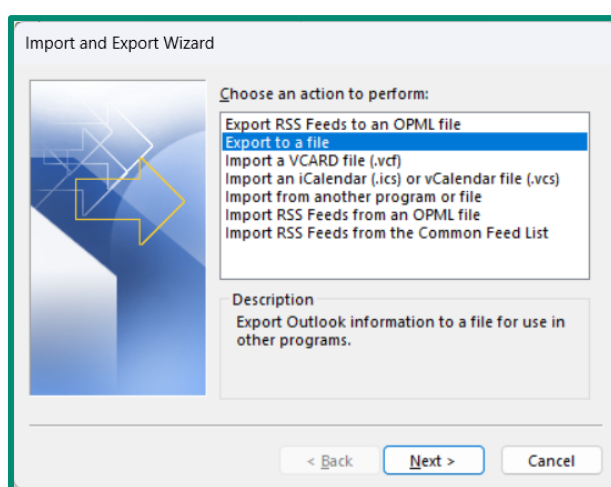
Introduction

The following is a step-by-step guide to exporting the contents of an email mailbox or selected folders from the Outlook desktop email client. It is not possible to download emails in bulk from the Outlook webmail interface, only via the desktop client or the system admin interface. The export will be in Microsoft's open but proprietary .PST format. In addition to emails, you will also have the option to include calendar data, contacts, and tasks within the export.

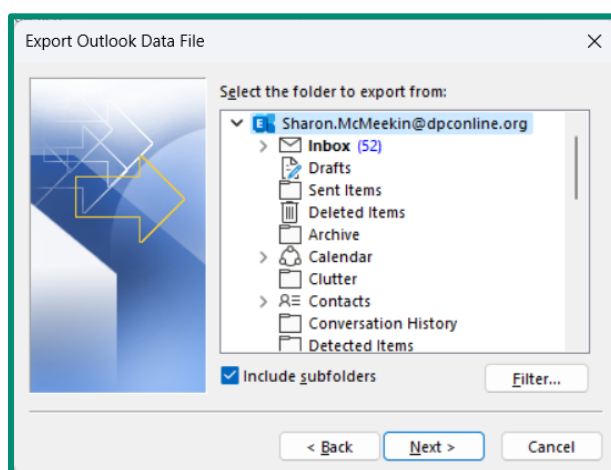
To carry out the next steps, the relevant account must be set-up and logged in on the Outlook desktop email client.

Step-By-Step Guide to Exporting Email from Outlook

- Click on the following menu items in turn:
 - File**
 - Open & Export**
 - Import/Export**
- An **Import and Export Wizard** window will appear. From this select **Export to a File** and click **Next**.



- Select **Outlook Data File (.pst)** from the “Export to a file” options and click **Next**.
- You will now be able to select what you wish to download. To download the entire mailbox click on the **account name** at the top of the list as shown in the example below. Alternatively, you can download an individual folder. For both options you have the choice of including or excluding subfolders using the **Include Subfolders** checkbox. Unfortunately, Outlook does not allow you to select multiple folders, only a whole mailbox or an individual folder.



5. Outlook provides a powerful set of filtering options that would allow you to undertake pre-ingest appraisal. These are available by clicking on the **Filter...** button. Details of the range of filter options available are included in the next section.
6. Once you are happy with your folder selection and filtering click **Next**.
7. Click on the **Browse** button and use the resulting file explorer dialogue to select where you would like to save the exported file and what you would like it to be called. Outlook allows for updates to existing export files, and the options provided on this screen refer to how duplicates should be handled during this process. If you are doing regular exports from an Outlook mailbox, you may wish to use this function to reduce duplication.
8. Click **Finish** to complete the selection process.
9. Outlook will now allow you to add a password to protect the exported content. This can be good practice if the exported content will be saved to a shared storage area. If adding a password, it is important to have a process for securely recording and sharing this information to ensure access is not lost due to a forgotten password. Leave the boxes blank if you do not wish to add a password. Outlook will then ask you to input the password to begin the export process.
10. Depending on the size of the mailbox or folders chosen for download the export may take anything from a few seconds to several minutes.
11. You will find the exported file in the location chosen.

Filtering Options

The filtering functionality offered by Outlook is accessed via the **Filter...** button mentioned above, and three tabs within the dialogue box that appears: **Messages**, **More Choices**, and **Advanced**. We will look at the options available via each of these tabs in turn.

Messages

The messages tab allows filtering according to the following:

- Keyword search within the subject field, the subject field and message body, or frequently used text fields. Unfortunately, there is no clear list of what those text fields are.
- The **From...** and/or **Sent To...** fields, allowing filtering by specific email addresses.
- Where the account holder is:
 - The only person in the To line
 - On the To line with other people
 - On the CC line with other people
- By a number of **Time** criteria such as date message was sent or received.

More Choices

The more choices tab allows filtering by a number of criteria relating to the message itself. These include:

- Categories assigned using Outlook's tagging functionality
- Whether or not the email has attachments
- The level of importance assigned to the email
- Flags that have been added using Outlook's flagging functionality
- The size of the email in KB

Advanced

The advanced tab allows the creation of specific conditional criteria on a field-by-field basis. With the hundreds of fields that can be included, this provides the opportunity for setting extremely detailed criteria. Multiple criteria can also be added to refine the filtering further.

Module 4.5: Capturing Email from Gmail

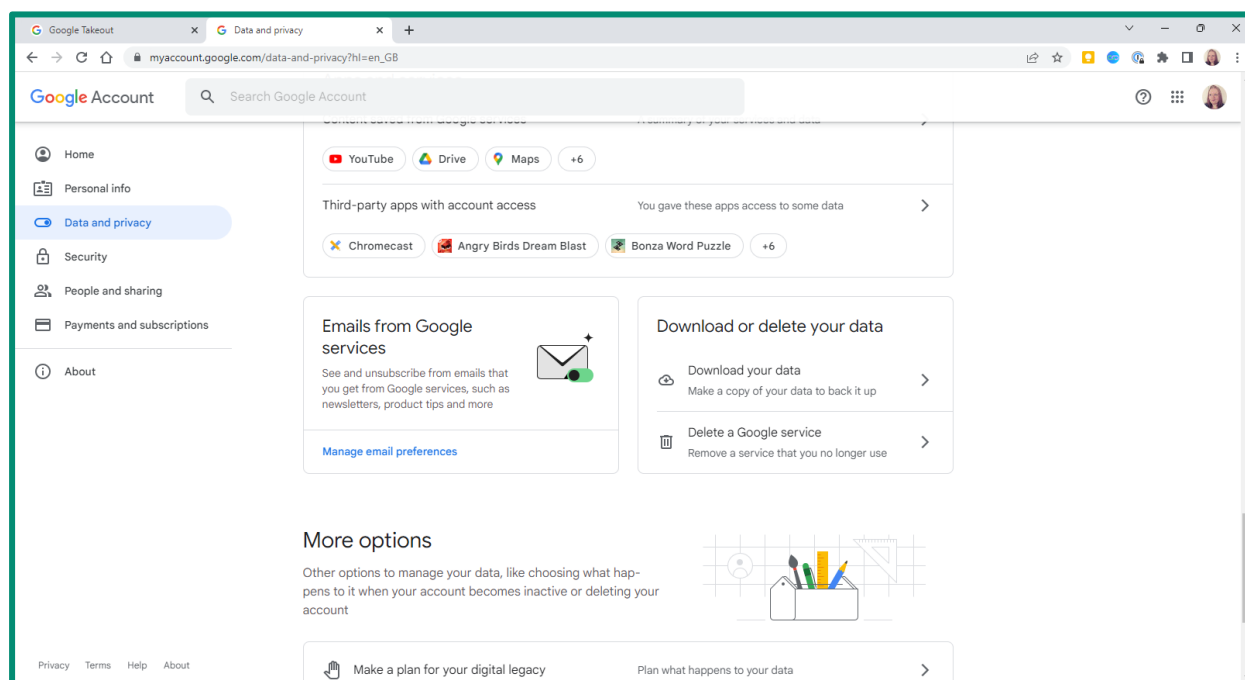
Introduction

The following is a step-by-step guide to exporting the contents of an email mailbox from the Gmail webmail service. The mailbox itself is exported in the MBOX format. Gmail also exports any custom user settings in a JSON file. In theory, importing the two files back into Gmail would recreate the original mailbox including any customizations and preferences the original user had set.

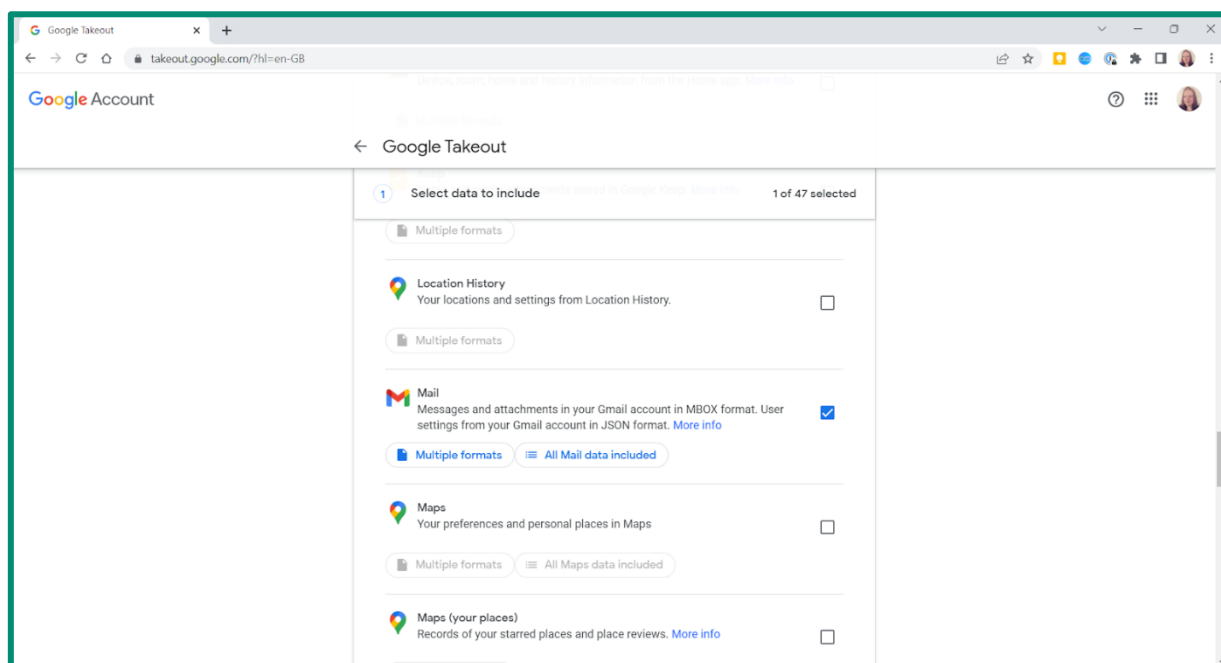
To carry out the steps below you must be logged in to the relevant Gmail account.

Step-By-Step Guide to Setting Up a Gmail Export

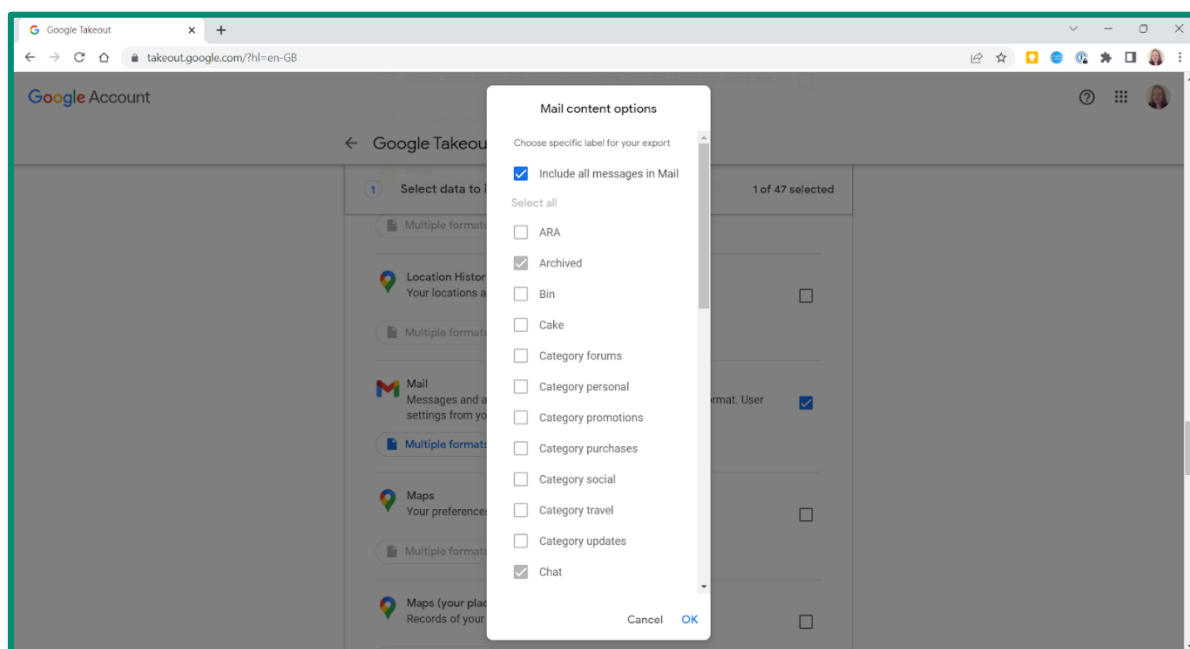
- 1) Go to the webpage takeout.google.com
 - a) Alternatively, you can select **Manage Your Google Account** by clicking on your profile picture at the top right of the screen while on any Google page. Then select **Data and Privacy**, scroll down to **Download and delete your data** under the **Data from apps and services that you use** heading, and select **Download your data**.



- 2) This page allows download of user data from a wide selection of Google Services. As we are only interested in Gmail at this time, start by clicking on **Deselect all**.
- 3) Scroll down to **Mail** and click to select



- 4) At this stage you can select which folders within the mailbox will be included. The default is all messages and folders. To change this click on **All Mail data included** and select the folders or categories of messages you wish to include and then click **OK**.



- 5) Scroll to the bottom of the page and click the **Next step** button.
- 6) The next section allows you to select the file type, frequency and destination for the export.
 - a) Under **Destination** the first option is **Send download link via email**. This will send a link to the inbox of the account you are logged in to. The link will allow you to download the exported mailbox when it is ready, saving it to a location of your choosing. Alternatively, you can choose to send the export directly to a folder in Dropbox, Google Drive (the same account or a different one that you are logged in to), OneDrive or Box using the relevant **Add to...** option. Choosing to export to Dropbox, OneDrive or Box will require linking your account for the relevant service to the Google account in use. The linking process will happen when you are ready to start the export process.
 - b) Under **Frequency** you can choose to **Export once** or set up a bi-monthly automated export for a year (6 exports).
 - c) Under **File type & size** you can select how your export will be compressed and packaged, as either a .zip file or a .tgz file. The .zip format produces a slightly larger file but is more widely supported, with most operating systems offering utilities for “unzipping” them.
 - d) Also under **File type & size**, you can specify the maximum size of the export. If the mailbox exceeds this size then it will be split across multiple MBOX files. Capturing the mailbox in a single file will make preservation more simple. You can see the amount of

storage currently being used by the account by visiting one.google.com/storage while logged in. *For info: free Google accounts include 15 GB of storage (this includes data stored across all Google apps including Drive), Google One subscriptions offer more storage, starting at 100GB.*

- 7) You are now ready to start the export process.
- a) If you have chosen to receive a download link via email you will be able to click the **Create export** button now. You will then see a message saying that “Google is creating a copy of files from Mail”, that this process may take a long time, and that you will receive an email once this is complete.
 - b) If you have chosen to save the export in Google Drive the **Create export** button will also be displayed. You may then be asked to re-enter the password for your account.
 - c) If you have chosen to save the export to a storage service other than Google Drive you will be presented with a **Link accounts and create export** button. This is where you will be asked to enter information to link your storage service.

When the Export is Complete

You will receive an email to notify you when the export process has been completed. For a small mailbox this can happen relatively quickly, even within minutes, for a large mailbox it may take a few hours to a few days.

If you selected to have a download link sent, this will be included in the notification email. The link will allow you to download the .zip or .tgz file to your local storage environment. If you selected to have the export saved to a storage service, there will be a link to the folder within the email. The .zip or .tgz file will have been saved in a folder titled **“Takeout”**.

At this stage you will be able to unzip the folder and access the MBOX and JSON (if available) files. You are now ready to process the MBOX file for preservation.

Module 4.6: Introduction to ePADD, Including Capture/Import of Email

Introduction

ePADD is an open-source application, available under an Apache Public License v. 2.0, that “supports archival processes around the appraisal, ingest, processing, discovery, and delivery or email archives”. Its development is managed by Stanford University’s Department of Special Collections and University Archives, in partnership with Harvard University and the University of Manchester.

It has been chosen for inclusion within this learning pathway due to the range of functionality it offers and because it is the most technologically accessible of the specialist email preservation

tools. ePADD is operated using an easy to open browser-based graphical user interface, whereas other tools require scripting and database development skills to be deployed. Remember, a list of tools and systems that can be used for email preservation is included within the learning pathways' accompanying resources.

Included across the remainder of the learning pathway are a number of tool demos of the functionality offered by ePADD. These are not intended to be fully comprehensive, but rather provide a baseline of competency for using the tool. A detailed [User Guide](#) is available from [the ePADD website](#). Also included within the Additional Resources is an example MBOX file and some suggested tasks you can use to help further familiarize yourself with ePADD.

In this module we will provide an overview of ePADD's email preservation capabilities, describe system requirements and installation, and show you how to import emails into the tool. This content has been produced using ePADD version 10.

ePADD Functionality

ePADD's functionality is split between four modules covering different elements of the email preservation lifecycle. They are as follows:

- **Appraisal** - This module provides users with functionality to gather and review an email archive. This includes methods to help determine the relevance and importance of email messages; to identify and flag sensitive messages; and to impose restrictions on access.
- **Processing** - This module enables the organization and editing of an email archive following the initial appraisal, ready for preservation. This includes the ability to add further restrictions, input metadata, and export various elements of the email archive including attachments.
- **Discovery** - This is a stand-alone module that facilitates web-based discovery of redacted email archives.
- **Delivery** - This module provides a platform for the provision of access to email archives within an onsite reading/search room setting.

To use each module of ePADD, the email archive must first have been processed through the previous modules in order. For example, to add an email archive to **Processing** it must first have been imported into and exported from the **Appraisal** module.

ePADD uses MBOX as its working format within the tool. It is possible to import other formats into ePADD and to also export email archives in the EML format, but these actions require the use of an integration with the commercial tool Emailchemy to power the conversion. The ePADD project team has negotiated a reduced license cost for Emailchemy for this use, more information is available [on the website](#).

System Requirements

Version 10 of ePADD can be run in 64-bit Windows, Mac, and Linux (Ubuntu) environments. The minimum required versions of each are follows:

- Windows 7 SP1/10
- Mac OS X 10.13/10.14
- Ubuntu 16.04

The computer in use will also need to have a minimum of 8GB of RAM.

As mentioned earlier, the application itself is browser-based and should be used with Chrome (68 or later) or Firefox (59 or later). You will also need to have a recent version of the [Java](#) runtime environment installed (11 or later).

Installing and Opening ePADD

ePADD can be downloaded from the [project's GitHub repository](#). Download the .exe file for use with Windows and the .dmg file for use on a Mac. It is worth saving this file somewhere easy to find on your computer as you will use it to open ePADD each time you access the tool. Opening the tool simply requires double clicking on this file, although once you close the tool you will need to restart your computer before you will be able to open it again.

There are a number of settings that can be altered according to personal preferences, such as where ePADD will store working files and which module the tool opens to by default (Appraisal in the first instance). The default for storage is in the local drive of the computer, for Windows this can be found at C:\Users\<username>\epadd-<module>. It is important to ensure before starting work with ePADD that you have sufficient storage space for the email archive.

We recommend that you do not alter these settings when you are first learning to use the tool, as such they are outside the scope of this learning pathway. Details of the settings that can be changed, and how, can be found in the [User Guide](#).

Importing or Capturing Emails

When first opening ePADD, the tool will default to the **Import Screen** of the **Appraisal Module**. This page (as shown below) provides a number of options for importing emails into ePADD from one or more files and/or capturing them directly from one or more email accounts. The ability to include multiple files or accounts allows you to import an entire account even if it has been split across multiple files (e.g. if it was larger than the Gmail limit for exported MBOX files) or to include multiple accounts that might belong to a single user.

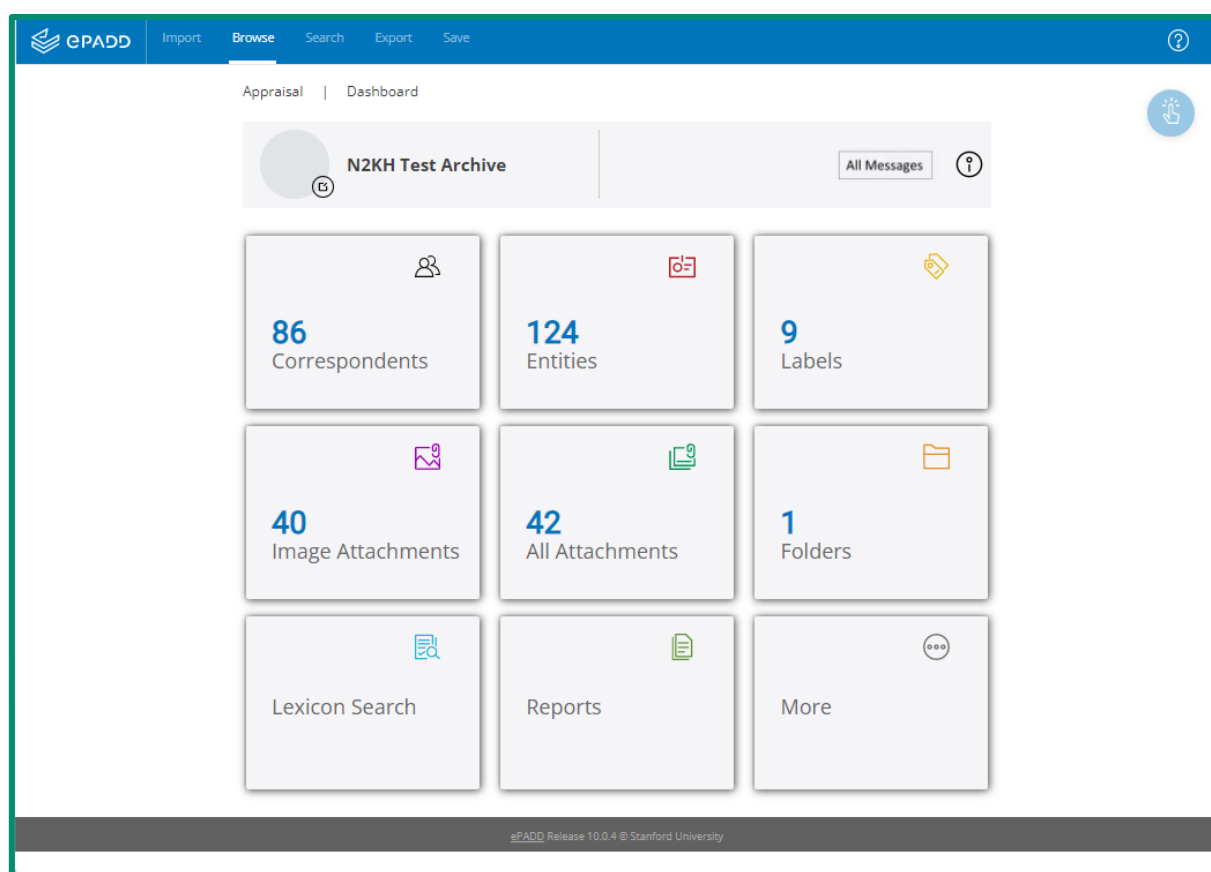
We will now look at each section in turn, describing their role when importing an email archive to ePADD.

1. **Help** - The question mark icon can be used to access general ePADD help and information. The pointing finger icon opens a prompt with help about the **Import Screen**.
2. **About this Archive** - This section allows you to import some basic metadata about the email archive: the name of the account holder, their primary email address, and the title you wish to give the archive.
3. **Public IMAP Email account** - This section facilitates capture of a complete mailbox direct from a webmail service. The login information for the account, email address and password, must be added here. Details of multiple accounts can be added for capture. Some webmail accounts will need to have IMAP connections “turned on” in the account settings to allow this process. *Note: due to an increase in Google security requirements for connecting apps, ePADD can not currently download directly from Gmail. This issue has been logged on the ePADD GitHub development site.*
4. **Private IMAP Email account** - This section can be used to capture directly from one or more privately hosted email accounts, such as the type provided by an employer. Here you must enter the location of the email IMAP server as well as login details. IT colleagues will be able to provide information on your organization’s IMAP server configuration.
5. **MBOX Files** - MBOX files exported from email accounts can be added here. This is the easiest way to add emails for import to ePADD. All that is required is to use the file browser to locate and select the relevant file and then enter a name for the source file to be included in the metadata.
6. **Non-MBOX email files** - This section allows the tool user to import email content that has been exported from accounts in formats other than MBOX. This functionality can only be used if the optional [Emailchemy add-on license](#) has been purchased as Emailchemy will be used to convert these files to MBOX as part of the import process. As with MBOX files you should use the file browser to identify the file to be imported and provide a name. You will also need to select the **Input File Format** from a list of possible originating email clients.
7. **Upload a Sidecar File** - This section allows the upload of additional documentation to accompany the email archive. This might include documents such as a depositor agreement/deed of gift or relevant license agreements. This functionality isn’t available during the initial import, but this screen can be accessed again later and these files added.
8. **Next** - Once the metadata and all email account or file information required has been added, clicking the **Next button** will start the import process.

55

The next screen will allow the user to select which folders from the email archives you wish to import to ePADD. Where an entire mailbox has been captured, this might allow a first round of appraisal. If you are confident in your understanding of what is included in each folder, you can choose to omit folders that contain only emails that are not to be retained. For example, you may choose not to include a “Personal” folder where the account holder has only stored non-work emails.

ePADD will then begin processing the email archive. This can take a considerable amount of time for large archives as ePADD is not only importing the email content into its internal database structure, but also carrying out various natural language processing and named entity recognition that will aid with appraisal and processing of the email archive. Once complete, the dashboard of the appraisal module will be displayed (see below). We will delve into the appraisal module as part of the next course.



Course 5: Appraisal and Processing

Module 5.1: Processing Email for Ingest

Introduction

Once emails have been captured from email or intermediary systems, they will need to be processed ready for preservation. Your processing workflow may include steps such as conversion to chosen preservation formats, (further) appraisal, sensitivity review, extraction of attachments, addressing embedded links, cataloging, and creation of preservation metadata. Tools will be an essential part of this processing and you may need to deploy a mix of open-source and commercial tools to carry out all of the necessary actions.

In this module we will look at three different approaches to processing, before focusing on format conversion issues, and handling attachments and different types of links. We will cover appraisal, sensitivity review, cataloging, and preservation metadata in future modules.

Three Approaches to Processing

Processing of email for preservation and access can involve a significant investment of resources, particularly staff time. With this in mind three approaches have developed, which require the investment of these resources at different stages in the preservation lifecycle. These approaches can be generally referred to as “Review”, “Restrict” and “Embargo”. The approach you decide to take may depend on the resources you have available, or you may take different approaches depending on the collection being processed.

Review

The first approach, Review, requires the biggest upfront investment of resources. Here, most, if not all, processing is carried out after capture of email and before ingest. This will include appraisal and sensitivity review, to ensure only those emails required for preservation are ingested. This process may be manual and/or semi-automated, and quality assurance will be an important element. As mentioned in the introduction, format conversion, addressing attachments and links, and generating descriptive and preservation metadata will also be required.

Restrict

The second approach, Restrict, pushes some of the processing actions to later in the life cycle, usually when an access request is received. In this approach the creation of catalog records and preservation metadata will still be required at the point between capture and ingest to allow for discovery of the preserved emails and proper management over time. Depending on your policies and preservation processes, you may also carry out format conversions and processing of attachments and links. The majority of appraisal, and sensitivity review and

redaction will be left until an access request is received. At this point appraisal and review of emails will only occur for those emails identified within the access request.

Embargo

The third approach can be referred to as “Embargo”. This is usually deployed at organizations who have mandated embargo periods before records can be accessed. Depending on the organization these may be short (e.g. 5 years) or last multiple decades (e.g. 50 years). In these cases, much of the processing may be pushed back until the embargo period has elapsed. As with the Restrict approach, there will be some requirements for the creation of metadata for preservation and discovery after capture. For other processing actions, you may wish to carry out a risk analysis to identify if other actions such as format conversion and processing of attachments and links will occur after receipt or will wait until after the embargo period. These decisions may need to be made on a case-by-case basis depending on the characteristics of the collection in question.

Format Conversion

The need for format conversion will depend on the capture methods used and decisions made in relation to preservation formats. In an ideal world, emails will be captured from the original email or intermediary system (e.g. an EDRMS) in a preservation ready format such as EML or MBOX. However, this may not always be possible, particularly when emails are received as part of an external deposit.

For files not in MBOX or EML formats, you may wish to undertake conversion to your preferred format for preservation. This format conversion is not included as standard in any of the open-source tools. It is, however, possible to buy a reduced-price license for Emailchemy for use as an integration with ePADD. Emailchemy and other tools such as Aid4Mail can also be used independently for conversion.

There has been little research to date on potential data loss during the conversion process, so it is important to employ robust quality assurance processes. If possible, within storage and other resource restraints, you may choose to keep the original format of the emails as well as the result of the conversion.

Attachments

The first decision that you will need to make in relation to attachments is whether they will be stored with the email content as MIME-encoded data, or if they will be held separately in their native format. The main benefit of retaining attachments as MIME data is that it retains the link with the original email. If the attachment is removed, you will need to devise another approach to managing links between attachments and their parent email.

Many would argue that removal of attachments so that they can be stored in their native format is the best option for preservation. This will allow preservation actions such as file

format conversion to be carried out in line with other preserved digital content of the same format. If you do opt for removing attachments, you will need to establish a persistent linking solution to maintain the relationship between attachment and email. This is usually done within preservation metadata: captured in a spreadsheet, database, or repository system, or included with a METS wrapper.

Also, some practitioners have experienced issues with emails being unintentionally deleted or corrupted during the process of removing attachments. Here again a robust approach to workflow testing and quality assurance is important.

Links to Files

In addition to files that are attached to emails, you will need to select an approach to content that is linked to from within the body of an email. Many organizations now have policies discouraging the use of attachments, as the many copies this can create use up a lot of storage space. Instead, users will include a link to a document held within a shared drive.

You will need to decide whether you wish to capture a copy of these files alongside the emails that reference them or if you will rely on them being captured as part of other preservation activities. At the moment, retrieval of these documents would likely be a time intensive manual process or a complex technical issue as none of the email preservation tools currently offer functionality to facilitate this process.

Web Links

Finally, you may also wish to develop an approach to managing web links within emails. Given the ever-changing nature of web pages, there is no guarantee that a web page linked to from an email will still be in existence or at the same address by the time an archived email is accessed. There are three options to address this issue:

0. Adopt a policy of treating web links in the same way as you would documents mentioned in traditional correspondence, i.e. pass the responsibility for finding that document on to whoever is accessing the preserved content.
1. Attempt to replace links within emails with links to the same pages within web archives.
2. As above but in addition, Archive the relevant web pages linked to.

Options 2 and 3 both involve complex technical requirements and processes that are not currently offered within existing email preservation tools. Option 3 would also require a significant resource investment, and possibly even a new preservation program. For these reasons most organizations will likely adopt option 1.

Module 5.2: Appraisal Decisions

Introduction

In the previous module we introduced activities that might be undertaken when processing captured email for preservation. In this module we will focus on one of those activities in more detail: Appraisal. It is worth bearing in mind that appraisal is particularly challenging for email compared with other content types because of the way email merges many kinds and sources of information.

Appraisal can happen at different stages in the email preservation lifecycle: pre-appraisal before capture, appraisal during processing for ingest, and, potentially, future appraisal of preserved emails as tools develop new functionality and/or your understanding of what constitutes an email record and how email collections will be used evolves.

Therefore, appraisal is rarely a one-time, static process, it is more likely that you will undertake a number of rounds of appraisal, re-appraising records and identifying them for permanent retention. Your approach to appraisal and when and how it happens will also be dependent on the tools, resources, and time you have available.

As part of the appraisal process, you may also be carrying out sensitivity review, identifying emails that contain personal or other sensitive or confidential data.

Appraisal Criteria

Before carrying out appraisal, it is important to understand exactly which constitutes a record for preservation. A Collecting Policy or Retention Schedule may contain high-level information on this, and should be one of the key documents consulted when carrying out appraisal. You may, however, have to make decisions about the practical application of this policy with regards to email.

Different collections may also require a different approach to defining the criteria for what constitutes a record. For example, when preserving the email of a senior executive, you will likely wish to preserve emails that relate to their work activities and any personal messages can (and perhaps should) be deleted. If you are appraising the collection an individual received as a deposit, the personal emails in the account might be some of the most important as they might provide evidence of relationships with other significant figures.

Developing a clear list of criteria for appraisal ahead of time will help facilitate the process, although you may need to adjust the criteria as you become familiar with the email collection. If there is the scope and time to discuss the contents of the email mailbox with the account owner, this can also help with defining criteria. In particular, they will be able to share details of potential categories of sensitive data within their email mailbox.

Some suggestions for categories of emails you may wish to address when carrying out email appraisal include:

- Personal emails
- Emails from mailing lists
- Spam and marketing emails
- Company circulars such as newsletters or all staff memos
- Emails covering personnel issues, e.g. payslips, contracts, sickness leave. These could be in relation to the account holder or another member of staff
- Sensitive or confidential data
- Emails from particular individuals or about particular topics

Appraisal Options

Email preservation tools offer increasingly useful and sophisticated functionality to aid in the appraisal process. Most tools include functionality for filtering by different criteria, searching, refining, and tagging emails.

For example, ePADD includes functionality to facilitate examination of a mailbox through a number of views that includes by correspondent, labels applied to the emails, attachments, and folders. A later module in this course will provide a demonstration of the appraisal functionality provided by ePADD.

Many email tools also provide options that will allow you to choose between deleting emails that should not or do not need to be retained, and redacting specific bits of information from emails. Redaction functionality can be particularly useful when dealing with sensitive or confidential information, it allows the retention of important emails whilst removing such information. This will allow access to emails that might otherwise need to be deleted or embargoed.

Natural Language Processing and Named Entity Recognition

Some of the most powerful functionality provided for appraisal by email preservation tools is facilitated by an approach called Natural Language Processing (NLP), and the sub-field Named Entity Recognition (NER).

NLP is an interdisciplinary field combining linguistics, computer science, and artificial intelligence. The goal is to develop computers capable of “understanding” the contents of digital content. This then allows the technology to extract information and key insights from the content, as well as organizing and categorizing them. As mentioned, NER is a subfield of NLP and it focuses on locating and extracting “named entities” within digital content and placing them into defined categories such as person names, organizations, locations, particular types of numbers like monetary values, quantities, times and dates, and many more.

Harnessing NLP and NER for email preservation can make a significant impact on the time and resources needed to appraise content. They can help to find patterns such as issues discussed and key correspondents, as well as aiding in the identification of potentially sensitive data. NLP

and NER functionality has been added to a number of email preservation tools such as ePADD and RATOM.

Using Forensics Tools for Appraisal

In addition to using tools specifically developed for email preservation as part of appraisal, some organizations have also included digital forensics tools as part of the workflow. Indeed, forensics tools offer a wide range of powerful functionality that can be deployed across the lifecycle of email preservation. To help facilitate appraisal, available functionality includes working with disk images, format conversions, and extracting “deleted” content.

The two most commonly used forensics tools are BitCurator and Forensic Toolkit (FTK). Instruction of the use of forensics tools is outside the scope of this learning pathway, but links to more information can be found in the resources that accompany the learning content.

Module 5.3: The ePADD Appraisal Module

Introduction

In this lesson, we will be looking at ePADD’s Appraisal module. This will include a description of the processes it can be used for, overviews of the main elements of the interface, and demos of tasks that can be undertaken.

Using the Appraisal Module

The ePADD Appraisal module provides users with functionality that will allow them to:

- Familiarize themselves with the imported email archive by using the various browse options to view information on correspondents, named entities, attachments, and more.
- Clean up any issues with the data, such as errors that have occurred during the import, and multiple entries for a single correspondent.
- Add metadata to the email archive
- Carry out a first round of appraisal

The [ePADD user guide](#) notes that users should build their own processes to meet the needs of their context and the email archives to be appraised, but recommend undertaking the following steps:

- 1) Review archive and clean data, including looking at:
 - a) Correspondents
 - b) Lexicons
 - c) Named Entities
 - d) Attachments
 - e) Labels

- 2) Add metadata
- 3) Appraise archive
- 4) Export the archive for use in the Processing module

Things to Remember

Before delving further into ePADD, there are a few useful tips to share to help with navigating through the tool and using its functionality. These are as follows:

- ePADD does not autosave your work, so it is important to regularly click the **Save** option from the main menu at the top of the screen.
- Navigation can be a little confusing at first as for some options ePADD will open the chosen screen in the same tab and other times will open it in a new tab. This means that sometimes returning to the previous screen requires using the back button, while other times you will need to close the current tab. It is common to find that you have multiple tabs open at once.
- During use of the appraisal module, ePADD will store all data in a folder at the following location by default:
 - **C:\Users\<username>\epadd-appraisal**
 - The email archive and accompanying files and metadata are stored here according to the [BagIt specification](#).
 - The data includes an unchanged version of the imported MBOX file (labeled the "Canonical" version by ePADD) and a version representing changes made during work using the Appraisal module ("Normalized" version).
 - Metadata for the email archive is saved here in both XML and JSON formats.

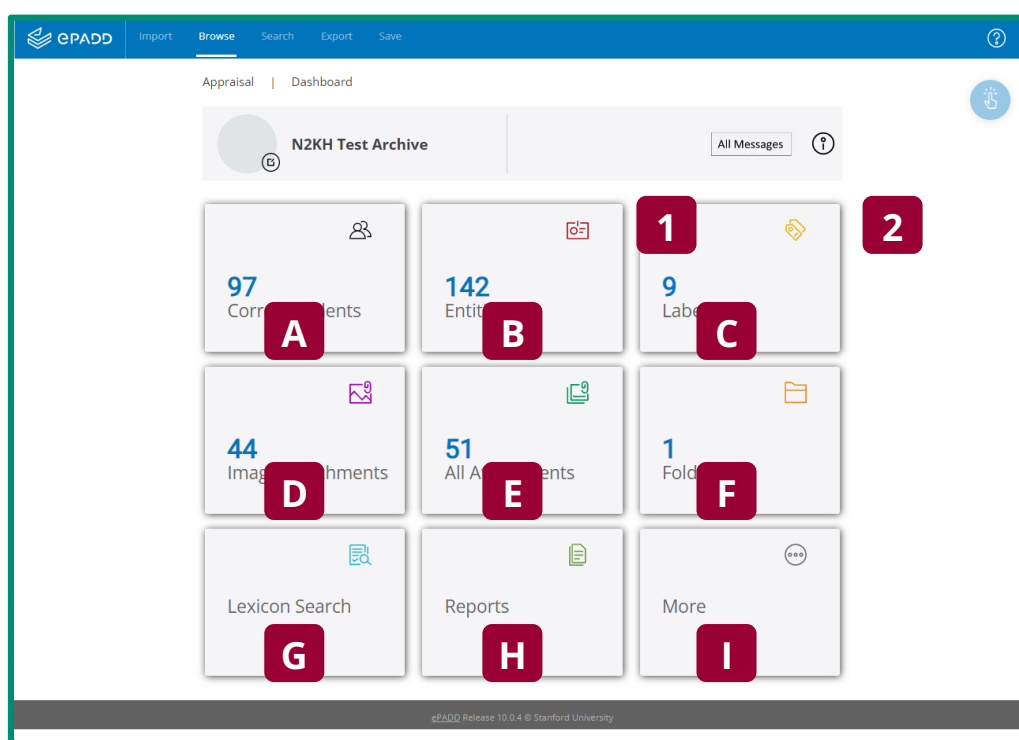
The Browse Dashboard

If one or more email mailboxes have already been imported into ePADD, the **Browse Dashboard** will be the first screen presented to the user after opening the tool (if not, the tool will default to the **Import** screen). This dashboard offers the user a wide range of different routes into browsing and interacting with the contents of the email archive. Options at the top of the archive information include the following:

1. Clicking on the **All Messages** button will allow you to browse through all of the messages within the email archive.
2. The ⓘ **Information** icon will allow you to access and edit information on the archive including:
 - a) High-level metadata.
 - b) Updating the checksum for the archive.
 - c) The option to delete the archive.

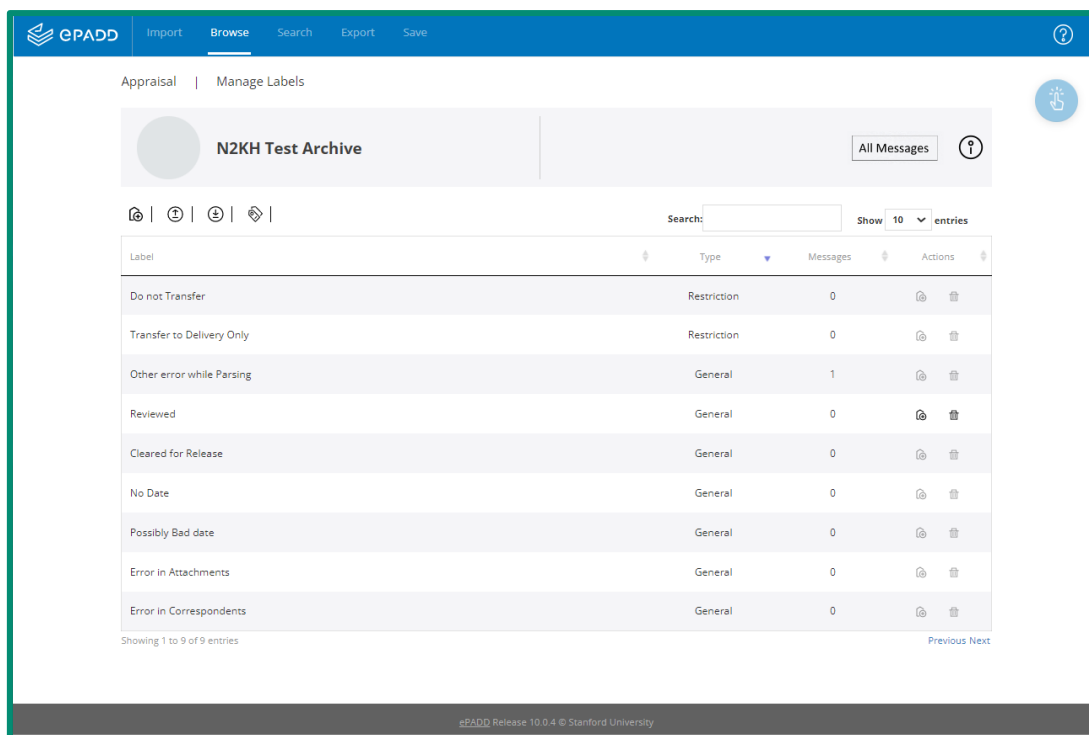
Below these options are a number of boxes that provide access to specific views on the contents of the archive. These are:

- Browse by a list of the **Correspondents** included in the messages
- Browse through a list of the named **Entities** (a word or phrase that refers to a real-world object, such as a person, location, organization, or product). These have been identified by ePADD using its NLP and NER functionality.
- Browse emails by the **Labels** that have been added to them. ePADD will have automatically added a label to any emails where errors occurred during import. Further labels will be added during appraisal and processing of the archive.
- Browse **Images** attached to the emails.
- Browse **All Attachments** included with the emails.
- Browse by the original **Folder** structure of the account(s).
- Browse according to available Lexicons (more on this later).
- View a summary **Report** of the archive, including information on duplicate attachments (this can help with deduplication during appraisal) and errors that occurred during import.
- Additional options under **More**. The functionality covered within this section of ePADD will not be covered by this learning pathway, but more details can be found in the ePADD User Guide.





















Depending on the contents of the email archive being processed, you may use some or all of the browsing options listed above. When starting to use the tool, it is useful to familiarize yourself with each.

Browsing a List



The screenshot shows the EPADD interface with the 'Manage Labels' page selected. The page title is 'N2KH Test Archive'. Below the title, there are icons for various actions and a search bar. The main content is a table with the following data:

| Label | Type | Messages | Actions |
|---------------------------|-------------|----------|---|
| Do not Transfer | Restriction | 0 |   |
| Transfer to Delivery Only | Restriction | 0 |   |
| Other error while Parsing | General | 1 |   |
| Reviewed | General | 0 |   |
| Cleared for Release | General | 0 |   |
| No Date | General | 0 |   |
| Possibly Bad date | General | 0 |   |
| Error in Attachments | General | 0 |   |
| Error in Correspondents | General | 0 |   |

Showing 1 to 9 of 9 entries

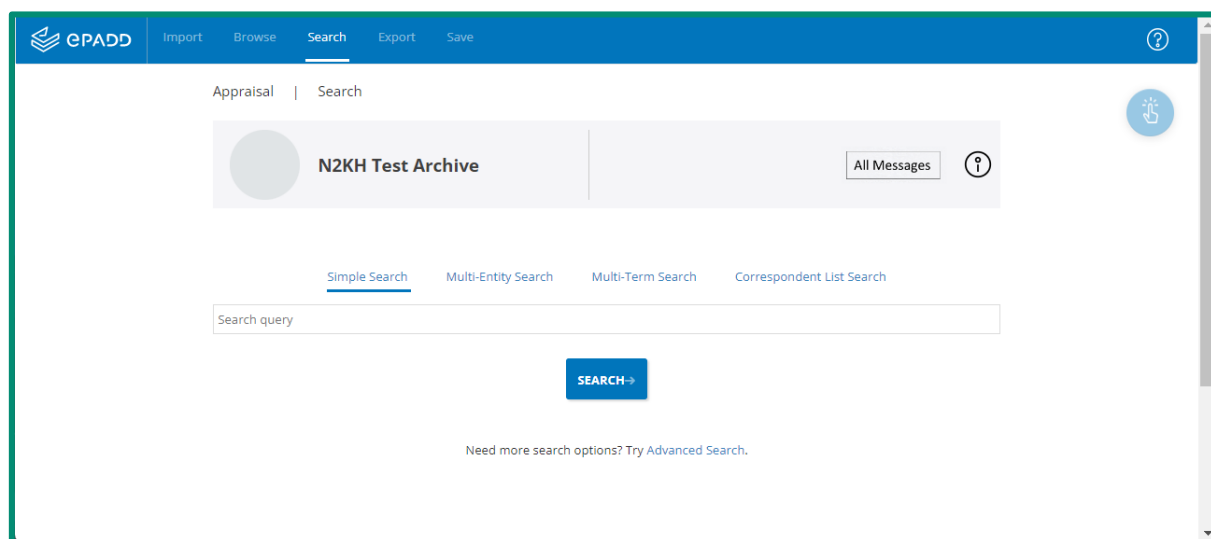
When you choose to browse by Correspondents, Entities, Labels, or Lexicon Search, you will be presented with a page showing a summary list according to the category selected. For example, if you choose to browse by Correspondents you will be presented with a list of the correspondents included within the archive and the number of sent and received emails they are included in.

The list page allows you to search within this list as well as reorder it according to the elements listed at the top by clicking on the appropriate column title. On the Entities list page you can order the entities alphabetically by Entity Type or by the number of entities identified in that category (e.g. Person, Place, Organization, etc.).

Depending on the page, clicking on a list item will either allow you to browse into a more granular entry within the category (for high-level lists under Lexicon and Entities) or it will take you to a page where you can browse the relevant messages (Correspondent and Labels lists and the lower-level entries of Lexicon and Entities).

Each page also provides a range of functions to allow you to manage, edit, import, and export data relating to these categories. These are accessible via the icons to the top left of the list. Each list page is slightly different, so we will not provide full details of the options here, but we will touch on how to use these options in the demos later in the module.

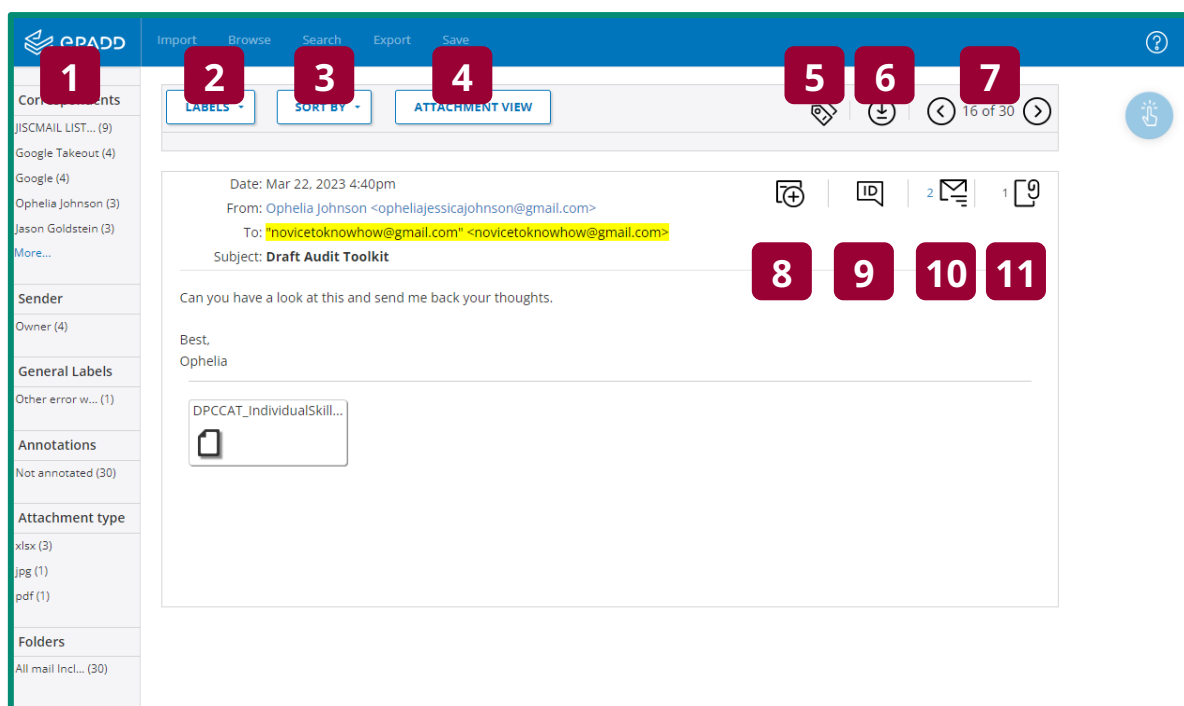
Searching



In addition to the range of browsing options, ePADD also offers a number of search functions to help identify groups of messages. The search options available are as follows:



- **Simple Search** - this allows searching of message headers, text, and attachments in document formats for the entered string of text.
- **Multi-Entity Search** - this allows searching for multiple entities at once. Each should be entered on a separate line. This will only return messages that contain entities identified by ePADD. Entities where results were found are highlighted. Messages can then be browsed by each term in turn.
- **Multi-Term Search** - similar to the multi-entity search, but using any keywords.
- **Correspondent List Search** - as with the two multi-searches but using email addresses within correspondent information.
- **Advanced Search** - The advanced search function allows searching across a wide range of specific criteria relating to message contents, elements, and the categories used by ePADD. A full list of the fields offered is outside the scope of this module.





Browsing Messages



The screenshot above shows the ePADD interface for browsing and working with individual messages, which can be reached through either the category browse options or searching. This example shows an example that has been reached through the Correspondents browse page, so the relevant correspondent is highlighted.

From **Browse Messages** page you can carry out the following actions:

1. Narrow your criteria further by using the facet list. For example, if you are viewing emails for a particular entity, you can then filter messages further by selecting a particular correspondent from the list of facets. ePADD will then display only the emails that mention the chosen entity **and** the selected correspondent.
2. Attach labels to individual emails, e.g. Add a "Do Not Transfer" label, using the **LABELS** button.
3. Sort the emails available by "Most relevant", "Newest first", and "Oldest first" using the **SORT BY** button.
4. View attachments for this group of messages by using the **ATTACHMENT VIEW** button.
5. Set labels for all messages in the group by clicking on the  **Label all these messages** icon. The resulting page allows the setting or removal of labels for all messages in the group.
6. Download all of the messages in this group by clicking on the  **Download messages as mbox** icon.

7. Browse back and forward through the group of messages using the arrow icons. The number of emails is shown between the arrows.
8. Add an annotation to the email by clicking on the  **Message annotation** icon, and entering text in the pop-up box. This can be a particularly useful function if multiple people are working on the archive.
9. View the ID number ePADD has assigned to the individual email by clicking on the  **Get message ID** icon.
10. View the thread that the message belongs to by clicking on the  **Open thread** icon.
11. Automatically scroll down to attachments on a long email by clicking on the  **Scroll down to attachments** icon.

The four key pages in the ePADD Appraisal module are **Search, Browse Dashboard, Lists, and Messages**. There are more pages you will find as you learn to work with the tool, but familiarizing yourself with the functionality mentioned above should assist with understanding how they can be used.

The Importance of Labels

The **Label** functionality in ePADD is the key method for actioning appraisal decisions as well as identifying messages that need to be embargoed, where issues or errors exist, and generally managing the appraisal process. Labels are split into two categories: restriction and general, the former to be used where emails will not be retained or there are access limitations, and the second for general management purposes.

Nine labels are preloaded in ePADD, five relating to error management and four to appraisal and access. The labels relating to errors are relatively self-explanatory, and the four relating to appraisal and access are as follows:


- **Do not transfer** - this label allows the identification of messages that should not be retained in the archival copy of the mailbox when it is exported from the current ePADD module.
- **Transfer to Delivery Only** - this will transfer only a redacted version of the email when exported for import into the Delivery module.
- **Reviewed** - marks that the message or messages have been reviewed.
- **Cleared for release** - allows you to clear for release any messages that have been previously embargoed.

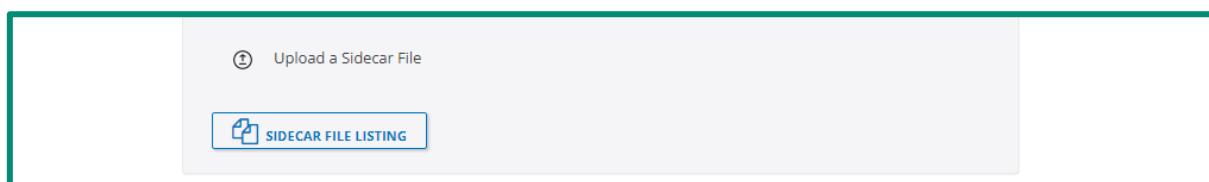
Additional labels can also be created for specific uses. For example, if messages are to be embargoed you can create a label that can be used to identify which messages are included and when the embargo will end. We will demonstrate this functionality later in the module.

Sidecar Files

ePADD also allows you to import supporting documentation for inclusion in the email archive, which are referred to in the tool as **Sidecar Files**. These sidecar files might include documentation such as any legal agreements relevant to the archive, e.g. depositor agreements, deeds of gift, or licenses allowing reuse of intellectual property.

Sidecar files can be added as follows:

1. Click on the **Import** item in the main menu.
2. Scroll down to the **Sidecar Files** section at the bottom of the import page.
3. Use the  **Upload a Sidecar File** icon to open a file explorer window.
4. Locate and select the relevant file and click **UPLOAD**.
5. A **Success** dialog box should display if the process has been completed.
6. Click the **SIDECAR FILE LISTING** button to view the list of files that have been added.
7. Repeat steps 3 and 4 to add additional files.



Exporting Content

The **Appraisal** and **Processing** modules of ePADD both offer a range of export options. We will examine these in detail when covering the Processing module. In the meantime, it is important to know how to export the email archive for use within the Processing module. This is completed using the following process:

1. Click on **Export** on the main menu.
2. Scroll down to the second box, **Export to next ePADD module**.
3. Use the file browser under **Specify Location** to select where you would like to save the exported content. Make sure this is somewhere easy to find and secure.
4. Enter a name for the content under **Exported Archive Name**. ePADD will have autofilled this box based on the Title you provided for the email archive. This can be edited according to your preferences or local naming conventions.
5. Click the **Export** button to begin the process. A "Success" dialog box will be displayed if the process has been completed correctly.

The email archive is now ready for import into the **Processing** module.

Module 5.4: Using the ePADD Appraisal Module


Working with Lexicons

It can sometimes be difficult to bridge the gap between a description of a tool and how to use it in practice. We hope to help with this process by providing examples of scenarios for using ePADD, and in the following three sections we will demonstrate how ePADD can be used for three specific tasks. The first task is using the Lexicon functions to identify sensitive data.


The ePADD Lexicon functionality allows you to identify messages that might include sensitive information based on sets of key terms that the tool will search for within the archive's messages. ePADD includes a number of preloaded Lexicon sets, but you will likely want to expand on these or add your own as you develop your email preservation program. When you edit a Lexicon set for an individual archive, you can then export that lexicon for use with future archives.

In our example, we know that the account owner often received emails about registrations for a training course that would likely contain personal data. To facilitate identification of these messages we'll add the term "registration" to the Sensitive Lexicon (the set of terms relating to sensitive data) and export the Lexicon for future use, before finding relevant emails and labeling them as "Do Not Transfer".

To add a term and update the new version of the lexicon:

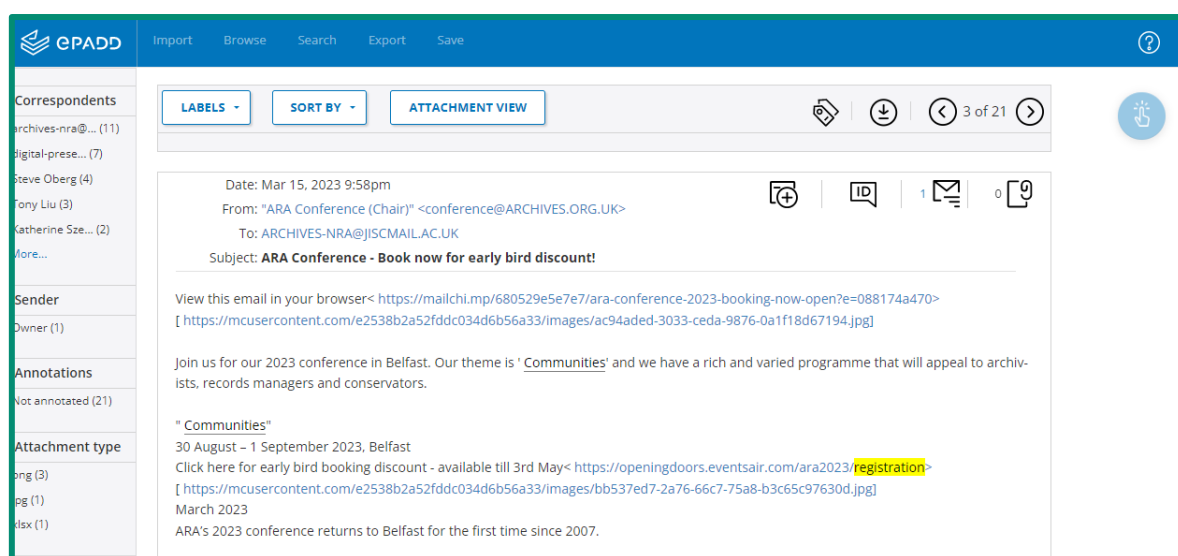
1. Click on **Lexicon Search** from the **Browse Dashboard**.
2. Click on the list item **sensitive**.
3. Click on the  **Edit lexicon** button.
4. Find the correct heading within the lexicon, in this case **Personally Identifiable Information**.



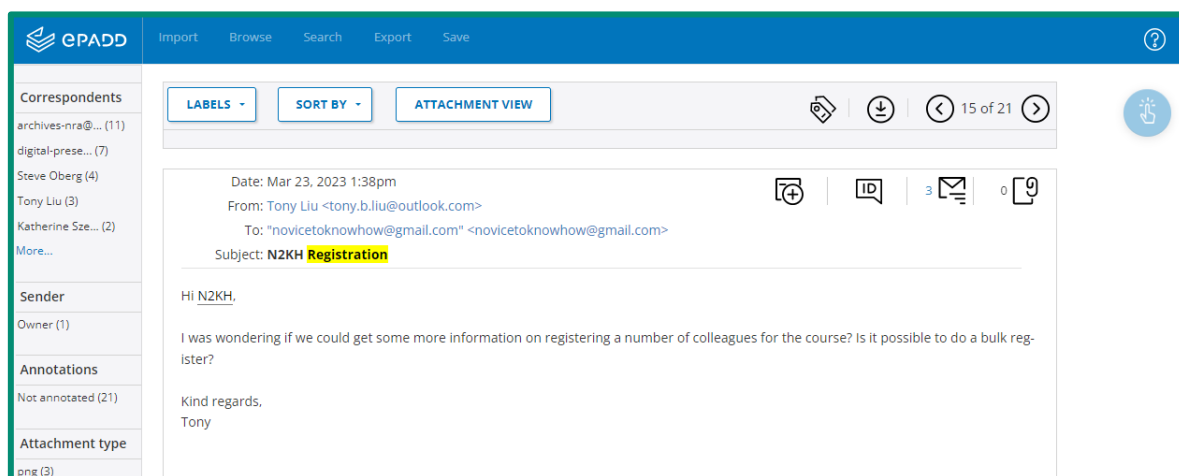
5. Enter the term in **registration** in the relevant text box. Terms should be separated using the vertical line | character and terms with multiple words should be included within double quotes, e.g. "registration details".
6. Click the **SAVE** button at the bottom of the page.
7. A **Success** dialog should be displayed. This can be closed.
8. Click on the  **Download Lexicon** icon to export the updated lexicon. This will allow it to be used with other archives.
9. Use the browser's back button to return to the lexicon list page.


To find and label emails with registration information using the updated lexicon:

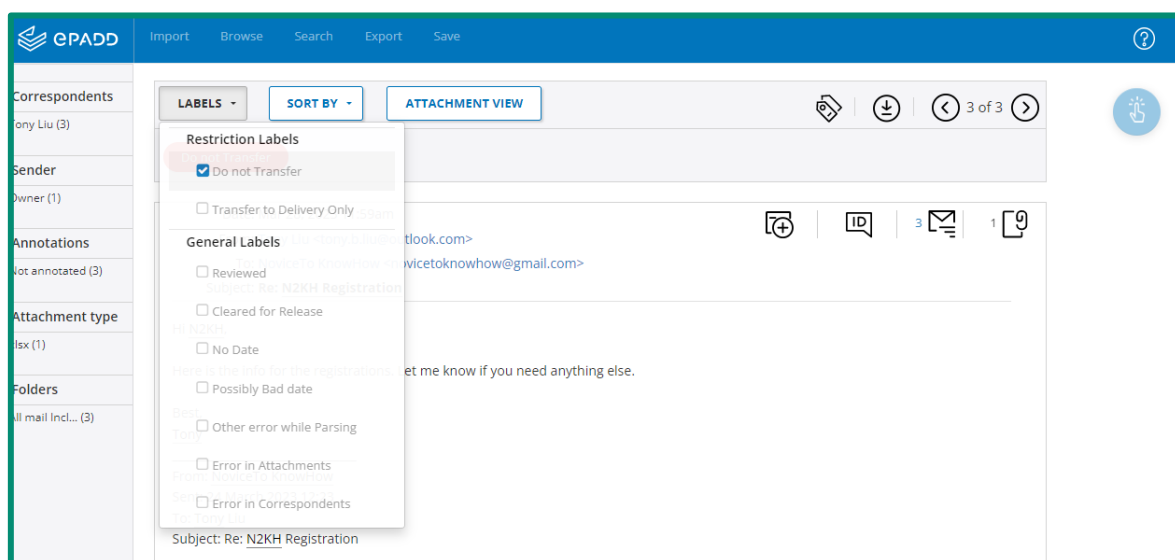
1. Click on **Personally Identifiable Information** with the lexicon list.
2. Use the browse arrows to find any emails with personal data. Lexicon terms will be highlighted in yellow.
3. A number of emails have been found that mention conference registration, so these can be quickly skipped past.

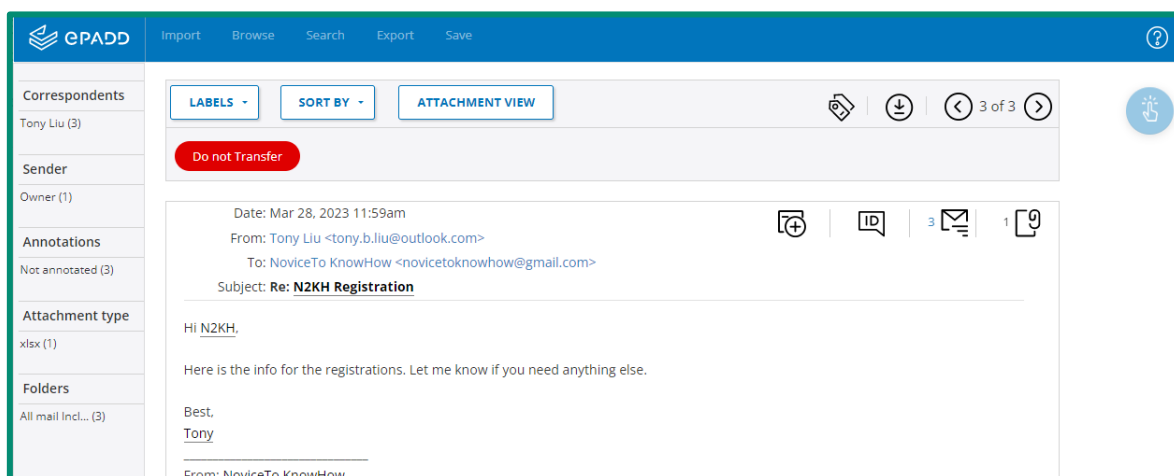


4. Email 15 includes a request to register colleagues for a course. Although it doesn't contain any sensitive data, it is part of a thread, so perhaps another email in the thread does.



- Click on the  **Open thread** icon to view the other related messages.
- Browsing through the messages shows the third includes an attachment that the message text indicates holds personal information. At this stage you may wish to download the attachment to confirm.
- Once confirmed, click on the **LABEL** button and select the **Do not transfer** option. This will add the label to the email. Now when the email archive is exported this message and attachment will not be included.
- Click **Save** in the top menu.




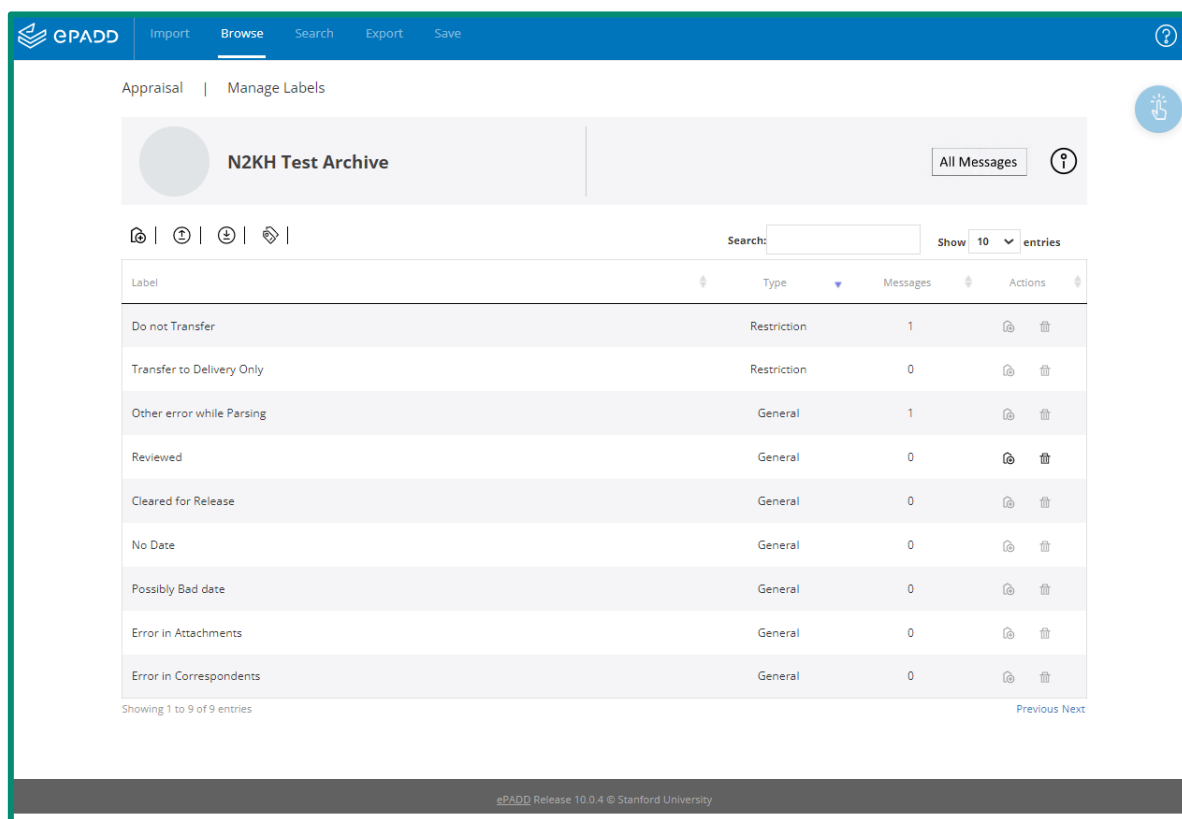


Working with Labels and Correspondents

In our next example we will look at how to create and action an embargo on a group of emails using labels. In this scenario, there is a ten year embargo on all emails from and to staff in a particular department. We have been provided with a list of relevant staff members from the department and we must locate their emails and apply the embargo.

To create the embargo label:

1. Click on the **Labels** box on the **Browse Dashboard**.
2. Click on the  **Create Label** icon to the top left of the label list.



Appraisal | Manage Labels

N2KH Test Archive

All Messages

Search: Show 10 entries

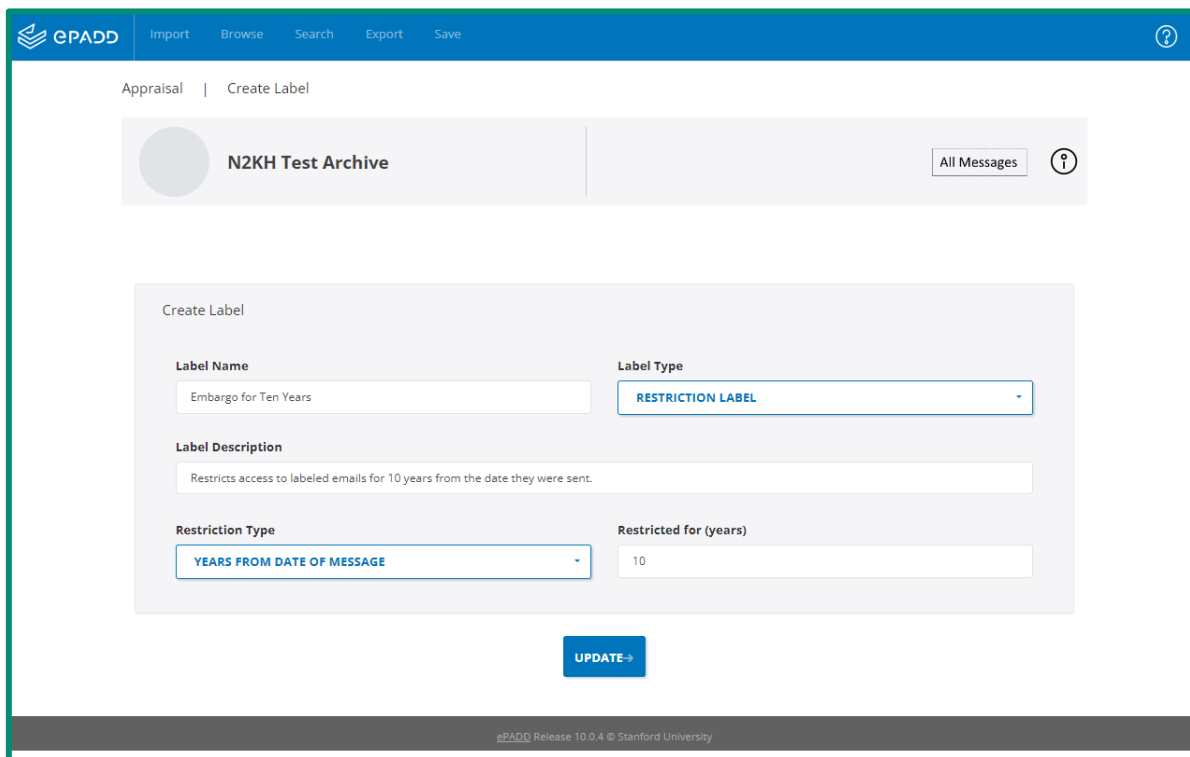
| Label | Type | Messages | Actions |
|---------------------------|-------------|----------|---------|
| Do not Transfer | Restriction | 1 | |
| Transfer to Delivery Only | Restriction | 0 | |
| Other error while Parsing | General | 1 | |
| Reviewed | General | 0 | |
| Cleared for Release | General | 0 | |
| No Date | General | 0 | |
| Possibly Bad date | General | 0 | |
| Error in Attachments | General | 0 | |
| Error in Correspondents | General | 0 | |

Showing 1 to 9 of 9 entries

Previous Next

ePADD Release 10.0.4 © Stanford University

3. Enter information on the label being created on the **Create Label**. Always aim to make this understandable for another user. In this example we will enter the following:
 - a) **Label Name** - Embargo for Ten Years.
 - b) **Label Type** - RESTRICTION LABEL (as we wish this label to indicate an access restriction.)
 - c) **Label Description** - Restricts access to labeled emails for 10 years from the date they were sent (explains the purpose of the label.)
 - d) **Restriction Type** - YEARS FROM DATE OF MESSAGE (other options allow setting until a particular date, or "not actionable" which will act more as a note than an active restriction.)
 - e) **Restricted for (years)** - 10 (enter a whole number of years.)



GPADD | Import | Browse | Search | Export | Save

Appraisal | Create Label

N2KH Test Archive | All Messages

Create Label

Label Name: Embargo for Ten Years

Label Type: RESTRICTION LABEL

Label Description: Restricts access to labeled emails for 10 years from the date they were sent.

Restriction Type: YEARS FROM DATE OF MESSAGE

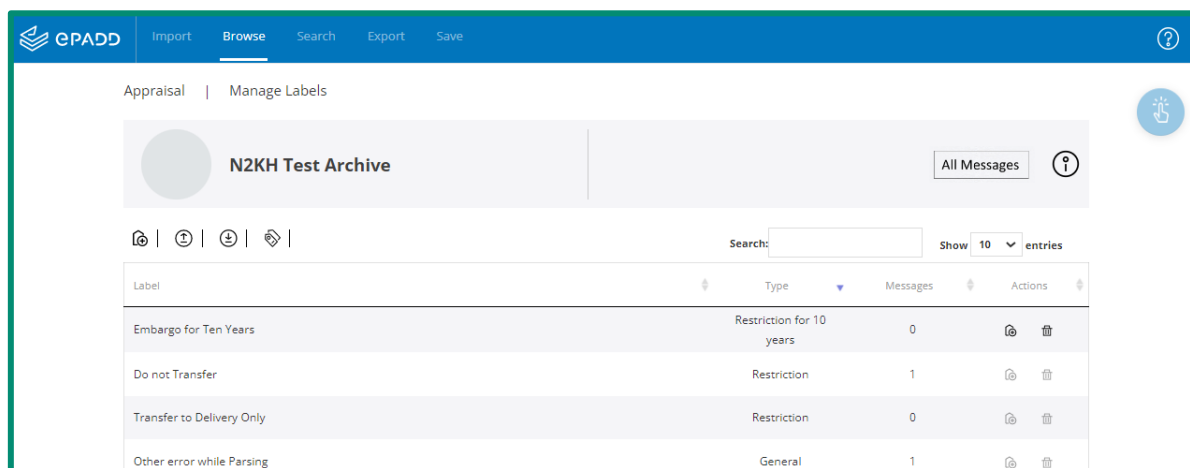
Restricted for (years): 10

UPDATE

GPADD Release 10.0.4 © Stanford University

4. Click the **Update** button to save the label.

A dialog box should let you know that the label has been successfully completed and the new label will now appear on the list along with the preloaded labels.



GPADD | Import | Browse | Search | Export | Save

Appraisal | Manage Labels

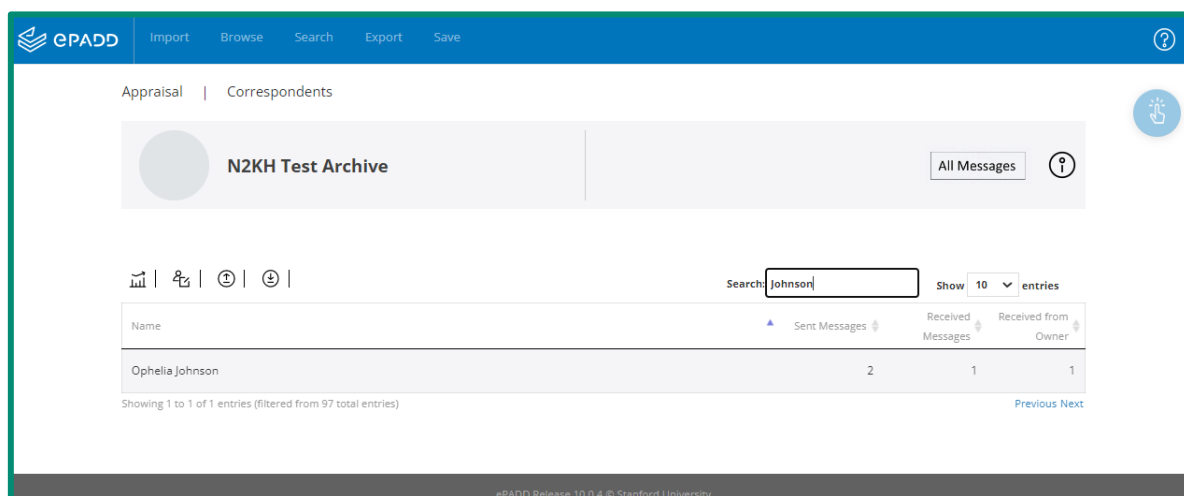
N2KH Test Archive | All Messages


Search: [] Show 10 entries

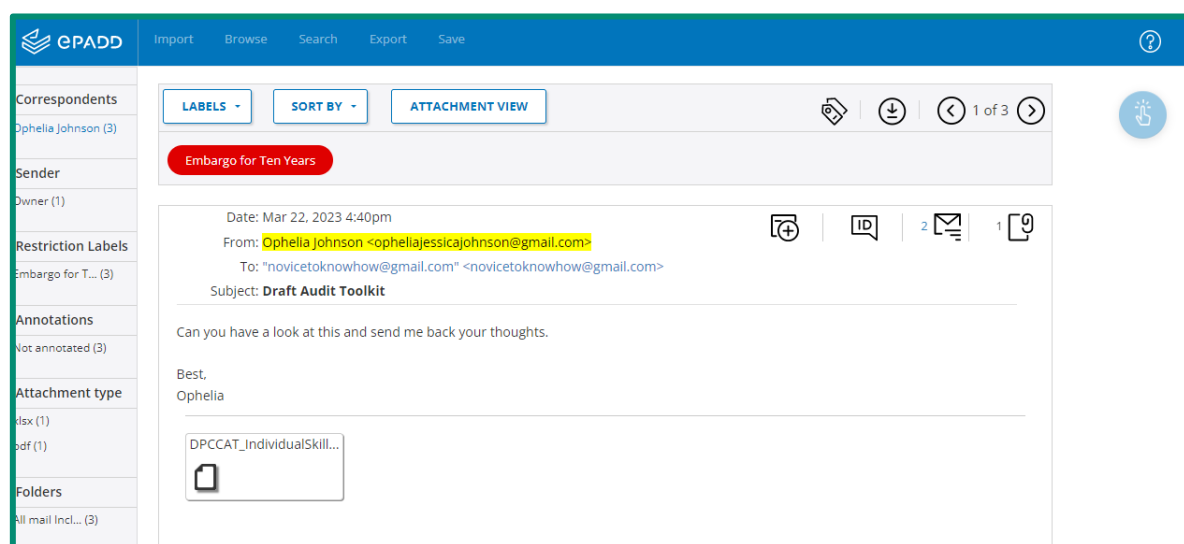
| Label | Type | Messages | Actions |
|---------------------------|--------------------------|----------|---------|
| Embargo for Ten Years | Restriction for 10 years | 0 | [] [] |
| Do not Transfer | Restriction | 1 | [] [] |
| Transfer to Delivery Only | Restriction | 0 | [] [] |
| Other error while Parsing | General | 1 | [] [] |

To label all of the emails to and from a particular correspondent:

1. Click on the **Correspondents** box from the **Browse Dashboard**. This will open a new **Correspondents** tab.
2. Locate the relevant correspondent by either browsing the list or searching for their name. Remember: you can organize the list by name by clicking on the relevant column header (it will order alphabetically by first name.) In this case we are looking for emails from Ophelia Johnson, so will enter Johnson in the search box.



- Click on the correspondent name to view all emails that include Ophelia Johnson. This will open a new **Browse** tab.
- Click on the  **Label all these messages** icon. This will open a new **Labels** tab.
- Click on the **Set for all** link for the **Embargo for Ten Years** label.
- Click **OK** on the dialog box to confirm you wish to set the label for all messages.
- Close the **Labels** tab and refresh the **Browse** tab and you should now be able to see the **Embargo for Ten Years** label attached to all emails.
- Click **Save** in the top menu.

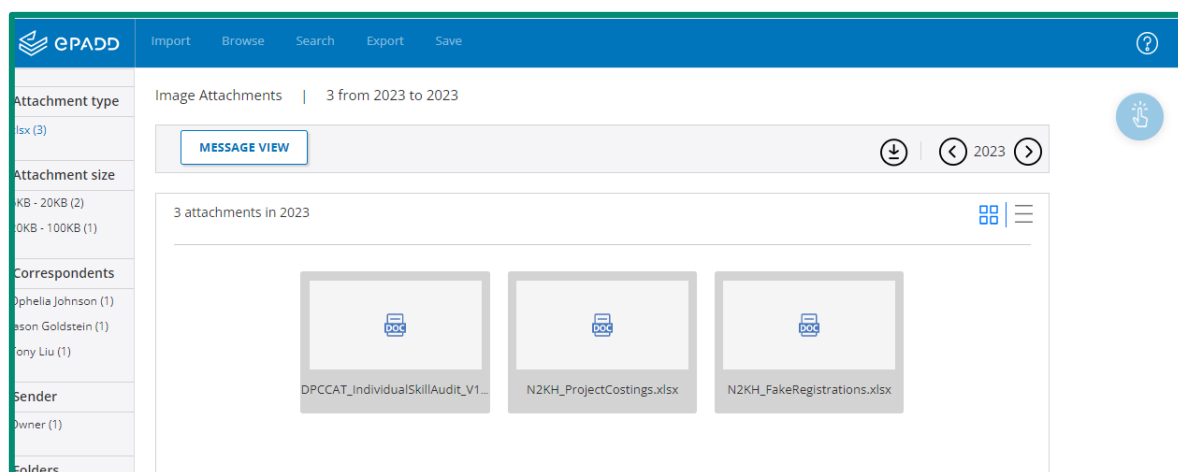


Working with Attachments

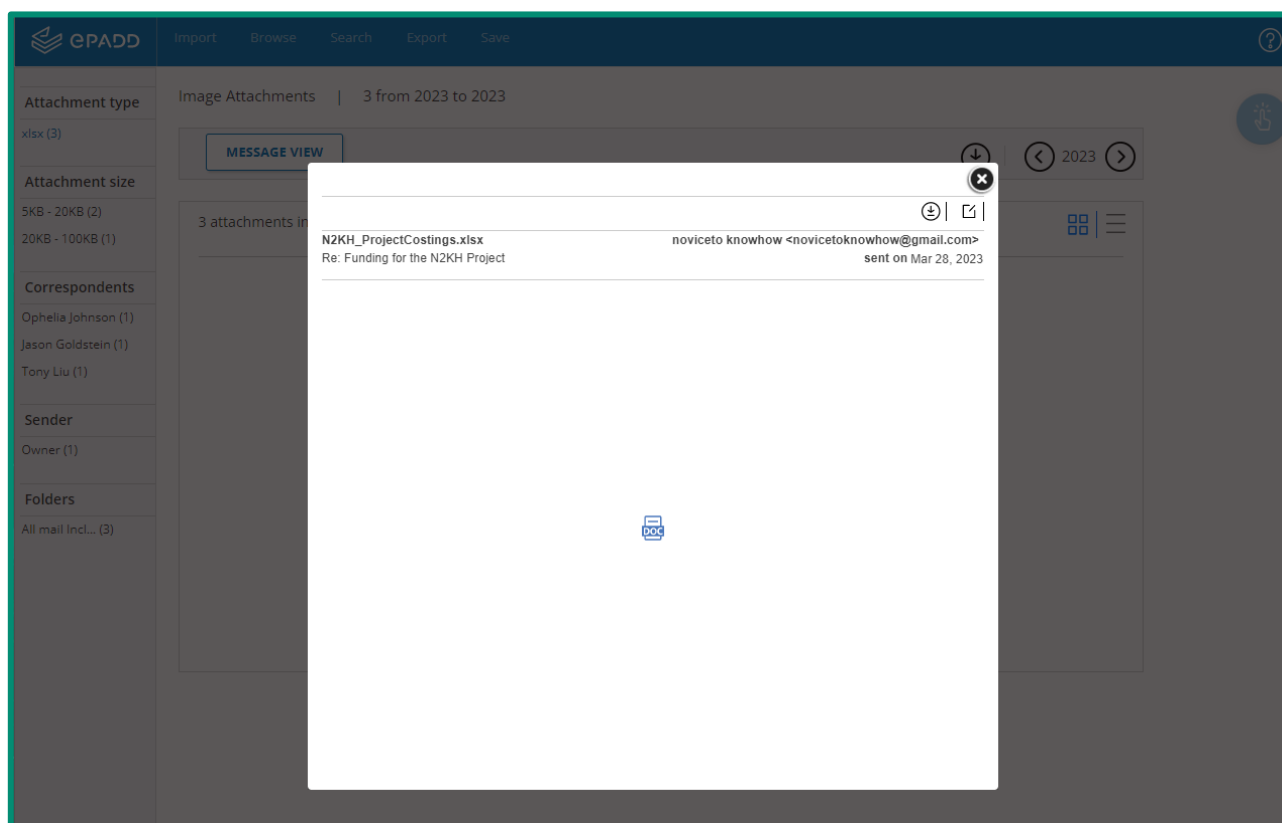
In this final example we will look at how to browse attachments for potentially sensitive information. In this scenario, the account user has told us they often received confidential budget estimates in spreadsheet formats, and according to the organization's retention plans this information should not be kept as part of the archive.


To find and identify these spreadsheets and label them correctly:

1. Click on the **All Attachments** from the **Browse Dashboard**. This will open a **Browse Messages** page that shows all attachments in a grid, this view can be switched to a list using the **List View** button.
2. Click on the **xlsx** item from the **Attachment type** facet list on the left. This will limit the attachments displayed to those in this format.



3. We can see here a spreadsheet called N2KHProjectCostings.xlsx. We might assume this will include budget estimates as described by the account holder, but we can download the file to double-check if needed. This is done by clicking on the **attachment info** and using the **Download attachment** icon in the dialog box that is displayed.



- Once we have ascertained if this does contain budget information we can click on the  **Go to message(s) associated with this attachment** icon and view the related message.
- Now we can add a **Do Not Transfer** label via the **LABELS** button, as described in the first example, so that this message will not be retained in the email archive.
- Click **Save** in the top menu.

Wrap-Up

These three examples offer an introduction to how ePADD could be used for appraisal. The original version of the mailbox used for these examples is included within the additional resources, so you can try out these tasks yourself. There is also a sheet with similar suggested tasks you can attempt to help you further familiarize yourself with ePADD.

Module 5.5: Capturing Metadata

Introduction

As part of the processing of email for preservation, it is important to develop and implement a plan for the capture of metadata to facilitate ongoing management, preservation, discovery of and access to email. As email is fundamentally a format built from a number of relatively well-

standardized and structured text elements, there is a lot of potential for automated, or semi-automated, capture of metadata.

In this module, we will look at key issues to be considered in planning and implementing your approach to capturing metadata, as well as examining key metadata elements that should be included.

Key Issues

Email header sections can contain a wealth of useful metadata. In particular, the transactional metadata generated to facilitate transmission of email has great potential for reuse. This will usually include information such as who created the email (the sender), who received the email (recipient(s)), when it was created, and if any attachments were included. Transactional metadata can also be useful in establishing the authenticity of emails by detailing how they were created and managed over time. It can also be useful for providing information to aid discovery and access. We will return to this use of header information and how it can help with cataloging in a later module.

Most email preservation tools include functionality for extracting metadata, with some offering the potential for automating the process. One issue to consider when selecting and using tools for metadata extraction is how the tool models metadata. Tools may capture metadata at the level of individual emails or at a higher level such as a folder or mailbox. It is important to ensure you select a tool that captures metadata at a level that is compatible with your plans for preservation formats and actions, as well as discovery and access.

You will also need to consider if you will capture metadata according to existing standards. Most, if not all, of the metadata you likely decide to capture for preservation will align with the schema defined by the PREMIS standard. Indeed many organizations have adopted the PREMIS standard for capturing metadata and paired it with the METS standard for structuring email information packages for preservation. It is important to remember, however, that PREMIS is a complex standard that can be intimidating and difficult to implement. Depending on the resources you have available and the scale of your email collections, it may be more appropriate to develop a simpler approach to metadata capture more suited to your context.

You will also need to consider how you will store your metadata. As with other types of digital content you might use any of a number of approaches that vary in complexity. The option chosen will depend on the resources you have available. Options include formats such as a spreadsheet, a database, XML files held with the content or separately, or within a repository or other asset management system.

Lastly, it is also always important to remember that due to security vulnerabilities inherent in email as a technology, there is a risk that the metadata held with emails may be altered and/or faked. You should consider the likelihood of this in relation to the collections you are

processing, particularly with regard to information that may be available on issues, such as spam scores.

Contextual Metadata

One of the key types of metadata it is essential to capture or create during the processing of email for preservation covers information relating to the context of how the email was created and structured, and about its provenance.

- Metadata elements you may wish to capture in this contextual metadata include:
- Details of folder structures within the email mailbox
- The relationships between emails and their attachments
- How the email mailbox relates to other preserved digital content
- Details of the provenance of the email, including information on the account holder, how and why it was selected and acquired, and who holds ownership of the content and any related IPR and other rights
- Information on any sensitivity issues, such as the inclusion of confidential messages or personal data

Preservation Metadata

You will also need to consider what preservation metadata you will need to capture or create to document all actions that are carried out on or in relation to the email for preservation. This will form an essential audit trail of all preservation actions. Preservation metadata will include, but is not limited to:

- Details of appraisal decisions and their implementation, e.g. details of emails that were deleted.
- Records of integrity checks, e.g. checksum information and an audit trail of when checks were carried out, the result, and any actions they triggered.
- Preservation actions that have been carried out, e.g. format conversions and quality assurance checks.

Provenance and Processing Metadata Model

The “Future of Email Archives” report includes a potential model for metadata to be captured about the provenance and processing of email for preservation. This might prove a useful starting point when determining what metadata you wish to capture within the context of your own organization. The model breaks down the metadata into key groups as follows:

Context of email creation

- Systems used to create email (e.g., web-based Gmail or MicrosoftOutlook client)
- Platforms used (computing environment)

- Type of use (business, personal, joint email account with family or organization, extracurricular)
- Who used the account (e.g., person or organizational unit comprised of persons and any changes over time)

Context of use or recordkeeping

- Account details
- Account name/username and legal/official name
- Length of time using the account
- System of arrangement (e.g., folders present or not and how they were used, whether the trash folder was used for drafts or for review)

Context of preservation or curation (many of these activities can be documented as PREMIS events)

- Who received the materials
- What acquisition processes were used to acquire the content
- What tools were used to inspect, appraise, inventory, review, and describe the materials
- Preservation policies that have been applied to the materials over time (when, by whom, for what purpose)

For personal papers or the donated records of an external organization, additional metadata may need to be tracked, such as:

- Selection criteria
- Email systems used by the donor
- Method by which an archivist captured the records

© Council on Library and Information Resources, 2018, CC BY-NC-SA

Course 6: Preservation

Module 6.1: Preservation Methods and Email

Introduction

The most commonly used preservation approaches that can be used for emails mirror those of other types of digital content: bitstream preservation, migration, and emulation. Though there are some email-specific issues to be considered when selecting your approach to preservation.

In this module we will highlight those issues to help you with decision-making around preservation approaches.

Bitstream Preservation

As with all types of digital content, bitstream preservation represents the baseline of preservation actions that it is recommended you undertake. If you have limited resources available and are not yet in a place where you can consider more complex preservation approaches, ensuring that you are preserving the bits may be your main goal. If this is the case, well-executed bitstream preservation will ensure you are well placed to undertake further preservation actions if capacity increases in the future.

Again, as with all digital content, bitstream preservation should focus on achieving the following:

- Maintaining multiple copies of the preserved emails (at least two copies, but three is often considered the best balance of risk and resource consumption)
- Ensuring that copies of the preserved emails are kept on different types of storage media, with at least one copy in a different geographical location.
- Carrying out integrity (fixity) checks whenever content is moved to ensure there has been no data loss.
- Carrying out integrity checks at regular intervals (e.g. every six months or yearly) to monitor for potential bit rot.

As mentioned previously in the learning pathway, a key decision you must make is whether to store attachments encoded as MIME along with their parent emails, or convert them to their original binary formats (e.g. DOCX or PDF). This decision will have an impact on your bitstream preservation actions as you will need to consider if attachments will be stored alongside email content or separately with content of a similar format. Storing attachments alongside email will help maintain their relationship, but maintaining them with similarly formatted content will likely create efficiencies if more advanced preservation actions are carried out.

File Format Conversion

Conversion of email from proprietary formats to the open IETF formats MBOX and EML has been mentioned earlier in the learning pathway. These formats are becoming *de facto* standards for email preservation, although some organizations also maintain email as XML or in Microsoft PST files.

This conversion process is often undertaken to facilitate long-term preservation, providing archives and other types of repositories with a homogenous collection. Use of MBOX and EML, in particular, alleviates reliance on any particular platform, as they can be imported into a wide range of email tools and clients for access.

Fewer tools exist for working with email as XML but the open “Email Account Schema” would help future interoperability and tools such as DArCMail are available.

Some organizations are also migrating emails to PDF for preservation, although this is generally only recommended for certain use cases, such as where the end users of the email archives have security concerns about loading another user's emails into their own email client. A specification for [“Using PDF to Package and Represent Email”](#) has been developed by the EA-PDF Working Group led by the University of Illinois at Urbana-Champaign.

Emulation

There has also been work looking at how to deploy emulation solutions for accessing preserved email by those looking to offer an “immersive experience”. Emulating an email client can be a particularly attractive access option given the inbuilt search and filtering functionality many offer. This would allow users to access the contents of an entire email mailbox and use that functionality to identify emails of interest, perhaps removing some of the burden of preparing content for access from those managing preserved content.

There are two main barriers to deploying emulation as a preservation option. The first relates to the known digital preservation issues relating to the programming skills and resources required for creating emulators, along with the added software preservation requirements. The second barrier is how to define what the original environment actually was, particularly in relation to current technologies where an account holder may use an email client, web-based client, and a mobile app at different times to access the same mailbox.

There are some interesting initiatives, however, that are looking at emulators for digital preservation purposes. For example, the [Emulation as a Service Infrastructure \(EaaSI\)](#) project includes work on emulators for past windows environments that would support access to preserved email.

To facilitate access to preserved email in its original environment using emulators you may wish to consider maintaining email in their original proprietary formats in addition to any preservation formats created through file format conversion.

Module 6.2: Designing an Archival Information Package for Email

Introduction

In previous modules, we have examined processing email for preservation, and decisions and actions around formats, metadata, appraisal, as well as options for managing attachments. Another important piece of the puzzle is determining how all of the files, metadata, and documentation will be organized and stored. This process is generally referred to as designing the Archival Information Package (AIP).

Your approach to the design of an AIP for email will be determined by a number of factors, including:

- Existing policy and processes for designing and implementing AIPs
- Where your AIPs will be stored, e.g. in a repository system, another type of digital asset management system, or in a preservation storage area
- Policies and decisions made around your approaches to preserving emails, e.g. will an AIP contain an entire mailbox, folder, or individual emails

In this module, we will highlight elements you may consider including in an email AIP, discuss how your approach to storage may influence AIP design, and offer a suggested directory structure for AIPs if you will not be storing them in a repository system.

What to Include

No matter how your AIPs will be stored, there are key elements that should be considered for inclusion. These include the email content itself, as well as all of the information that will be required to manage preservation over time.

Elements to consider for AIPs include:

- The email content in the “original” capture format, even if this is a proprietary format
- The email content in the normalized preservation format, if this process has been carried out
- Attachments if they have been separated from their parent emails
- Any additional information captured from the email account, e.g. details of tagging schemes used or subject and sender logs
- Documentation relating to the acquisition of the email content, e.g. depositor agreements/deeds of gift, IPR licenses, any supporting contextual documentation
- The metadata required to manage preservation (as described in module 5.4 - Capturing Metadata), e.g. fixity values, logs of any actions carried out

You may also wish to hold dissemination copies of the email content, and accompanying metadata, within the information packages or alongside them.

Storing Your AIPs

Designing your AIPs will also require consideration of how and where they will be stored. This might be in a repository system or other digital asset management system, or within a dedicated preservation storage area.

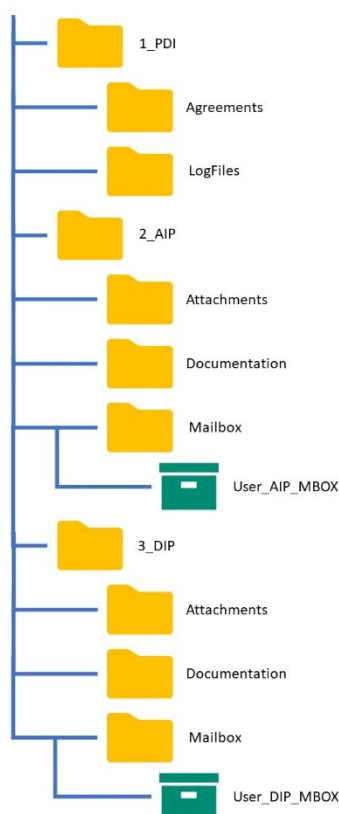
If you have a repository system in place, you will be able to capture most, if not all, of the information within the repository. In this case you will need to ensure the repository can accommodate all of the elements you wish to capture. How to ingest email into specific repository systems is outside the scope of this course, but the following issues should be considered.

Preparation for ingesting email into your repository may include establishing a new information package structure for email and/or updating existing settings. You will need to establish how the repository handles versioning, particularly if you will be maintaining an “original” version of the email as well as a preservation copy. Also, you will also need to know if and how the repository can store separate attachment files, and how the relationship between these and their parent emails can be retained.

Most repository systems will allow the capture of detailed preservation metadata, usually in line with the PREMIS standard, and will support capture of key documentation. If this is not included in the functionality, you will need to develop a plan for how and where to manage this separately.

If you do not have a repository system, you will need to develop a clear plan and structure for your AIPs using systems and resources you have available. There may be some scope for capturing elements of the documentation and metadata in other systems such as collection management and/or cataloging systems. For the storage of other metadata, you will need to consider if it will be:

- Held centrally, perhaps in a database or spreadsheet, with metadata relating to other preserved digital content;
- Held in a resource specifically for email collections, or;
- Maintained alongside the email content itself, perhaps in one or more spreadsheets or XML files.



The email content itself, accompanying attachments, and any metadata and documentation not captured in other systems, should be organized and saved to a storage area allocated for preservation. Devising a folder structure that can be replicated consistently for all email deposits is to be advised, and we will describe one such suggested structure in the next section. You may also consider documenting this structure using the METS metadata standard.

Suggested Directory Structure

In his “Preserving Email” Technology Watch Report, Chris Prom shares a suggested folder structure for email AIPs that was originally developed by the Smithsonian for use with their DArCMail tool. The structure defines three top level folders and subfolders as follows (and is shown in the illustration below):

- **A folder for Preservation Description Information (PDI):** this folder would hold all the necessary metadata and documentation relating to the email content, e.g. deposit agreements/deeds of

gifts, correspondence, log files, etc. Additional subfolders could be defined depending on the information you will capture.

- **A folder for Archival Copies of the Content:** this folder would hold all copies of email content and relevant attachments, including “original” versions and any converted preservation copies. Subfolders may be organized by version and/or content type (e.g. attachments, mailboxes, individual emails)
- **A folder for Dissemination Copies of the Content:** this folder would potentially hold copies of the email content prepared for access purposes. These would likely include one or more copies of the content that includes a subset of the emails from the account and/or content that has been redacted for privacy or confidentiality reasons. Metadata and the documentation users would require to be able to use the material, would also be included in this folder.

Exactly how the folders would be named and structured would depend on your own organizational context and local policies and procedures. It is important to state again that consistency in naming and structure is advised to make management and retrieval more efficient.

Module 6.3: The ePADD Processing Module

Introduction

In this lesson we will be looking at ePADD’s Processing module. This will include how to open the module and import an email archive, descriptions of the tasks it can be used for, and demos of tasks that can be undertaken.

Purpose of the Processing Module

The main purpose of the Processing Module, as described by the ePADD User Guide, is to transform an email archive accession into an email archive collection. To facilitate this transformation, the Processing module includes similar functionality to the Appraisal module to allow for further refining of the email content included, whilst also offering functionality for adding metadata and documentation to enable preservation and providing access.

The functionality in the Processing module allows:

- Adding and editing metadata required for a collection-level description and according to the EAD and PREMIS standards
- Adding labels to emails that relate to access levels
- Exporting the email archive in formats for preservation and access, as well as options to export individual elements

The Processing module also allows reconciliation of correspondent names with authority records from sources and services such as the [Library of Congress Subject Headings and Named Authorities](#), [OCLC Fast](#), [VIAF](#), and [Wikipedia](#). We will not demonstrate this functionality

as part of this learning pathway, but it is relatively easy to use and is covered in the [ePADD User Guide](#).

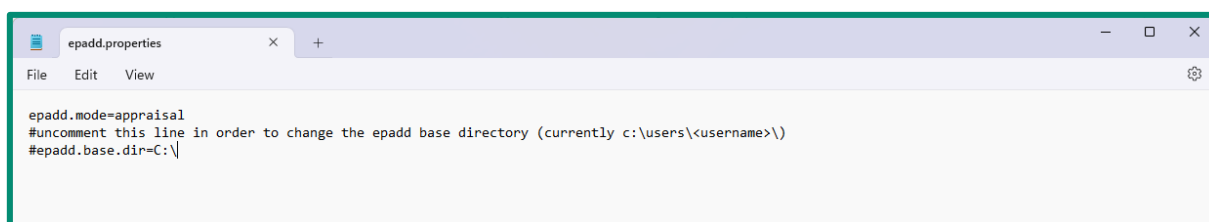
It is also important to note that you can add multiple email archives to the Processing module at once. You can switch between the collections in the tool by clicking on the **Collection** item in the main menu at the top of the screen.

Opening the Processing Module

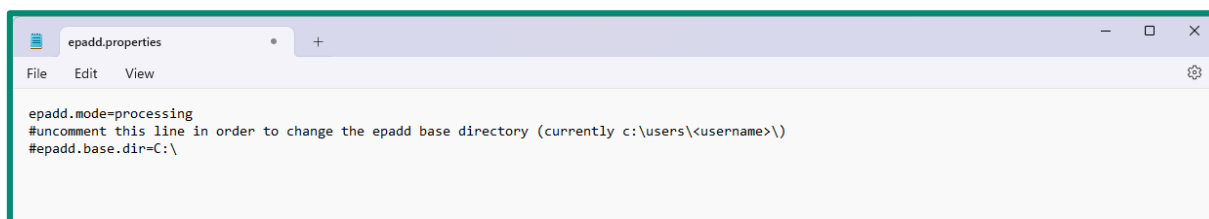
Unfortunately, there is not a simple button or menu item to allow switching between ePADD modules. Instead you must edit a “properties” text file to indicate the ePADD module you would like to open when you double-click on the ePADD.exe file.

Follow these steps to open the Processing module:

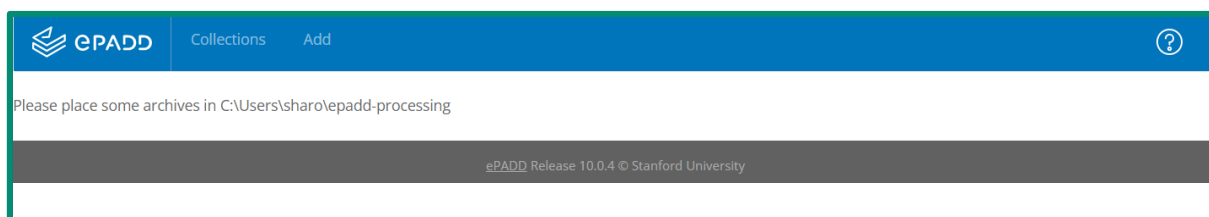
1. Use a File Explorer window to locate the **epadd.properties** file which will be saved in C:\Users\<username>.
2. Open the **epadd.properties** file. If you need to select an application to open the file, select **Notepad** (though any text editor will work). The file should look as shown in the image below.



3. Change the word **appraisal** to **processing** (see below)



4. Save and close the file.
5. Now double click on the **epadd.exe** file to open ePADD. The window that opens should look like the following.



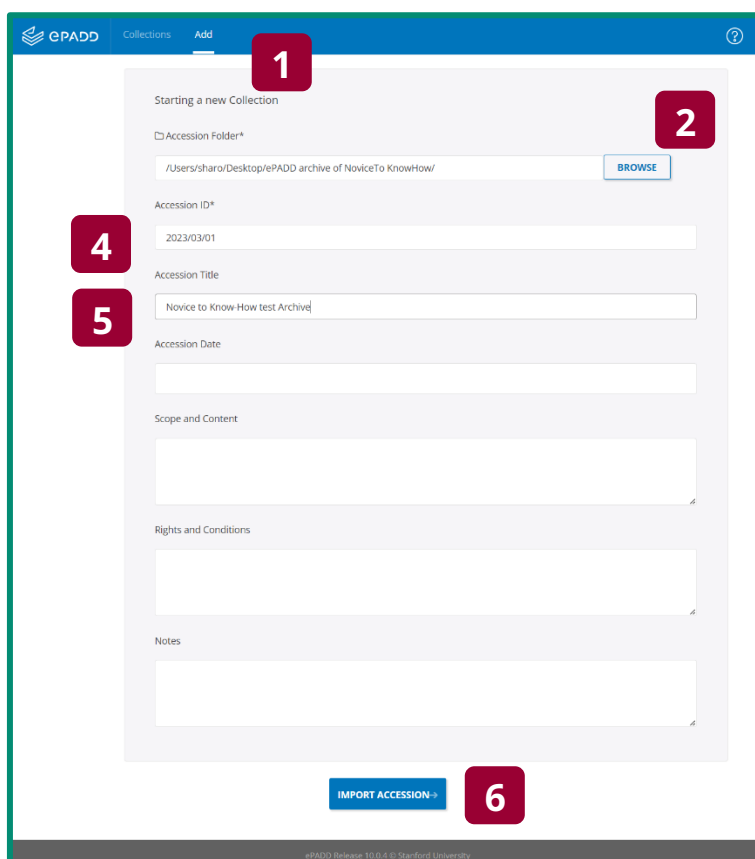
The same steps should be followed to open the Discovery and Delivery modules, replacing the current module name in the properties file for the desired module.

Adding an Email Archive for Processing

We are now ready to import an email archive into the Processing module. As part of this task it is possible to add elements of a collection-level description. We will, however, skip those elements during import to focus on demonstrating other functionality of the Processing module. A demonstration of how to add and edit this metadata after the email archive has been imported will be provided in the next section.

To import an email archive, the following steps should be followed:

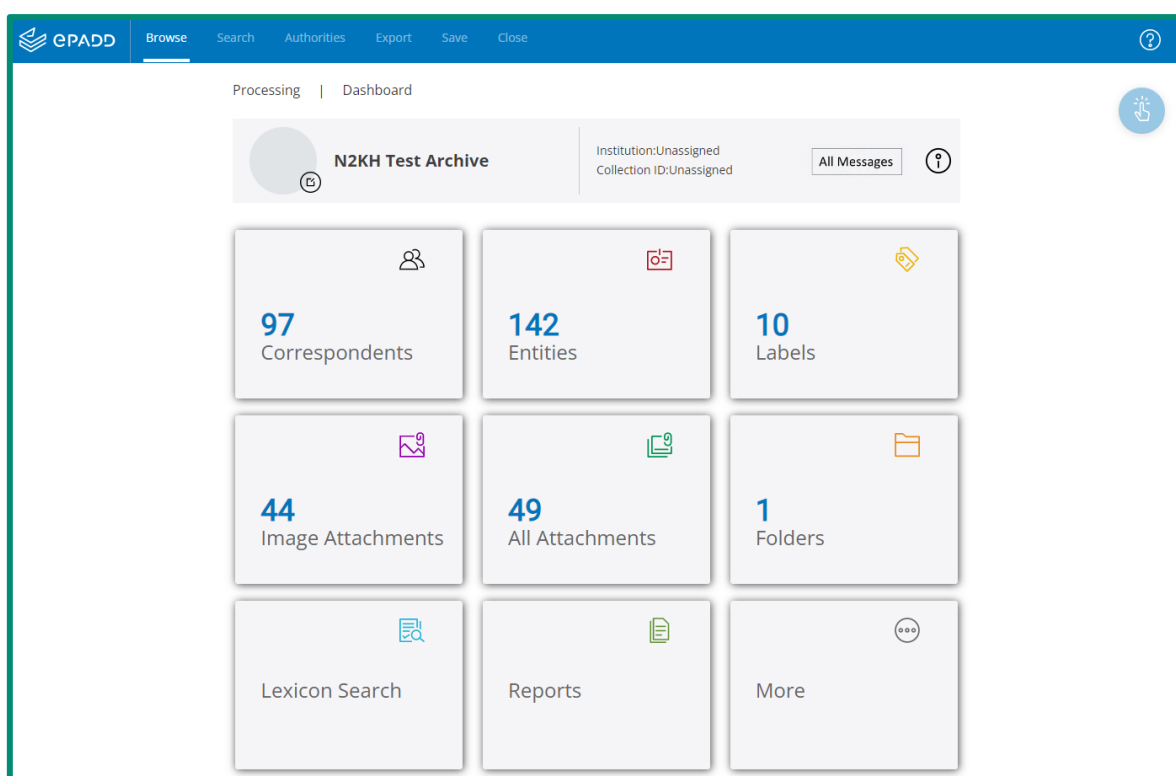
1. Click on the **Add** option from the main menu at the top of the page.
2. Click on the **BROWSE** button under **Starting a new Collection** to launch a file browser window.
3. Locate and select the folder that was exported from the **Appraisal** module and click **CONFIRM**.
4. Add an **Accession ID** according to your local conventions. This is the only piece of metadata that you are required to provide during import.
5. Add an **Accession Title**. This is not required, but it will make identifying the collection easier.



The screenshot shows the 'Starting a new Collection' form in the ePADD interface. The form is titled 'Starting a new Collection' and has a 'BROWSE' button next to the 'Accession Folder*' field. The form includes fields for 'Accession ID*', 'Accession Title', 'Accession Date', 'Scope and Content', 'Rights and Conditions', and 'Notes'. The 'Accession ID*' field contains the value '2023/03/01'. The 'Accession Title' field contains the value 'Novice to Know-How test Archive'. The 'Accession Date' field is empty. The 'Scope and Content', 'Rights and Conditions', and 'Notes' fields are also empty. The form is framed by a blue header bar with the ePADD logo and a 'Collections' tab. A red box with the number '1' points to the 'Add' button in the top navigation bar. A red box with the number '2' points to the 'BROWSE' button. A red box with the number '4' points to the 'Accession ID*' field. A red box with the number '5' points to the 'Accession Title' field. A red box with the number '6' points to the 'IMPORT ACCESSION' button at the bottom of the form.

6. Click the **IMPORT ACCESSION** button to begin the import.



If the import is successful a **Success** dialog box will be displayed. When you close this box you will find yourself at the **Browse Dashboard** (as shown below). The main part of this dashboard is identical to the one included in the **Appraisal** module, but there are additional items included in the main menu at the top of the screen.

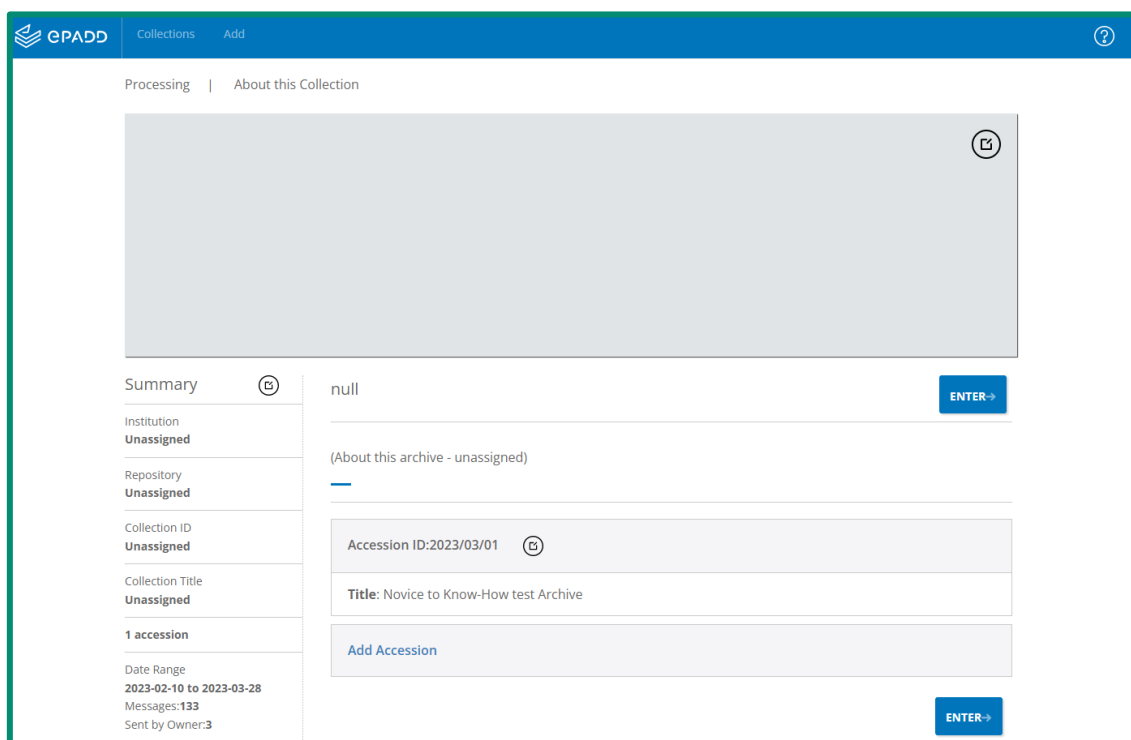


Editing Metadata

As mentioned earlier in this lesson, ePADD provides functionality to allow the capture of metadata for a collection-level description, as well as offering a range of fields that comply with the EAD and PREMIS standards. The following range of metadata fields are available using the entry method described below.

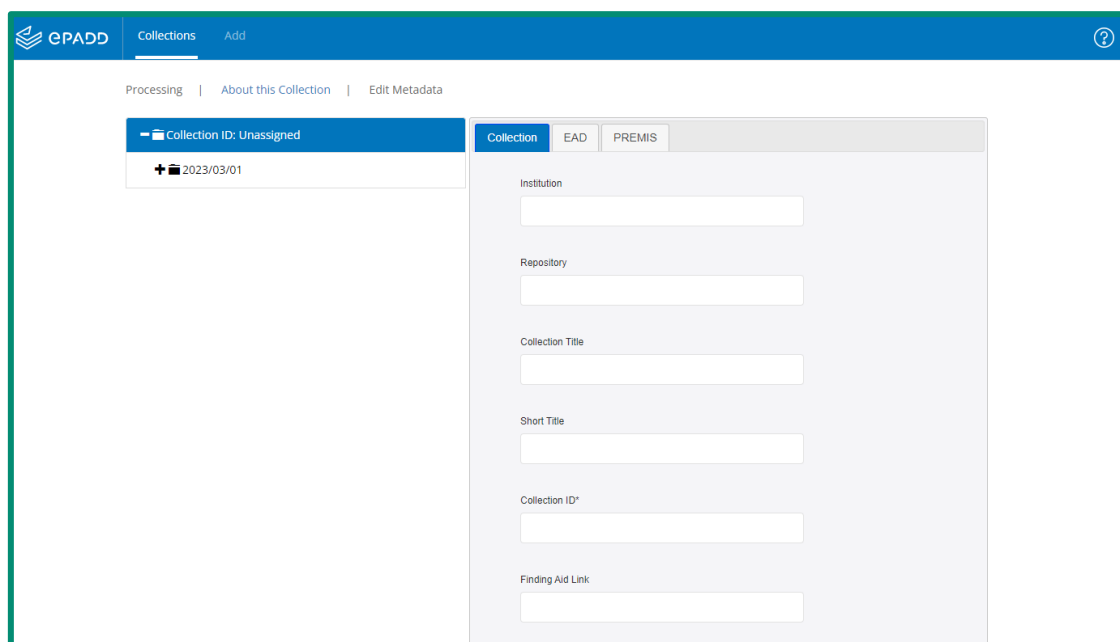
To enter metadata follow these steps:

- 1) Click on the **Collection** item on the main menu.
- 2) Click on the box relating to the desired email archive.
 - a) A collection profile picture can also be added from this page by clicking on the  **Edit** icon at the top right of the collection box.
- 3) Click on the  **Edit** icon next to **Summary** on the left of the **About this Collection** page.




4) Use the **Edit Metadata** page to enter the desired metadata.

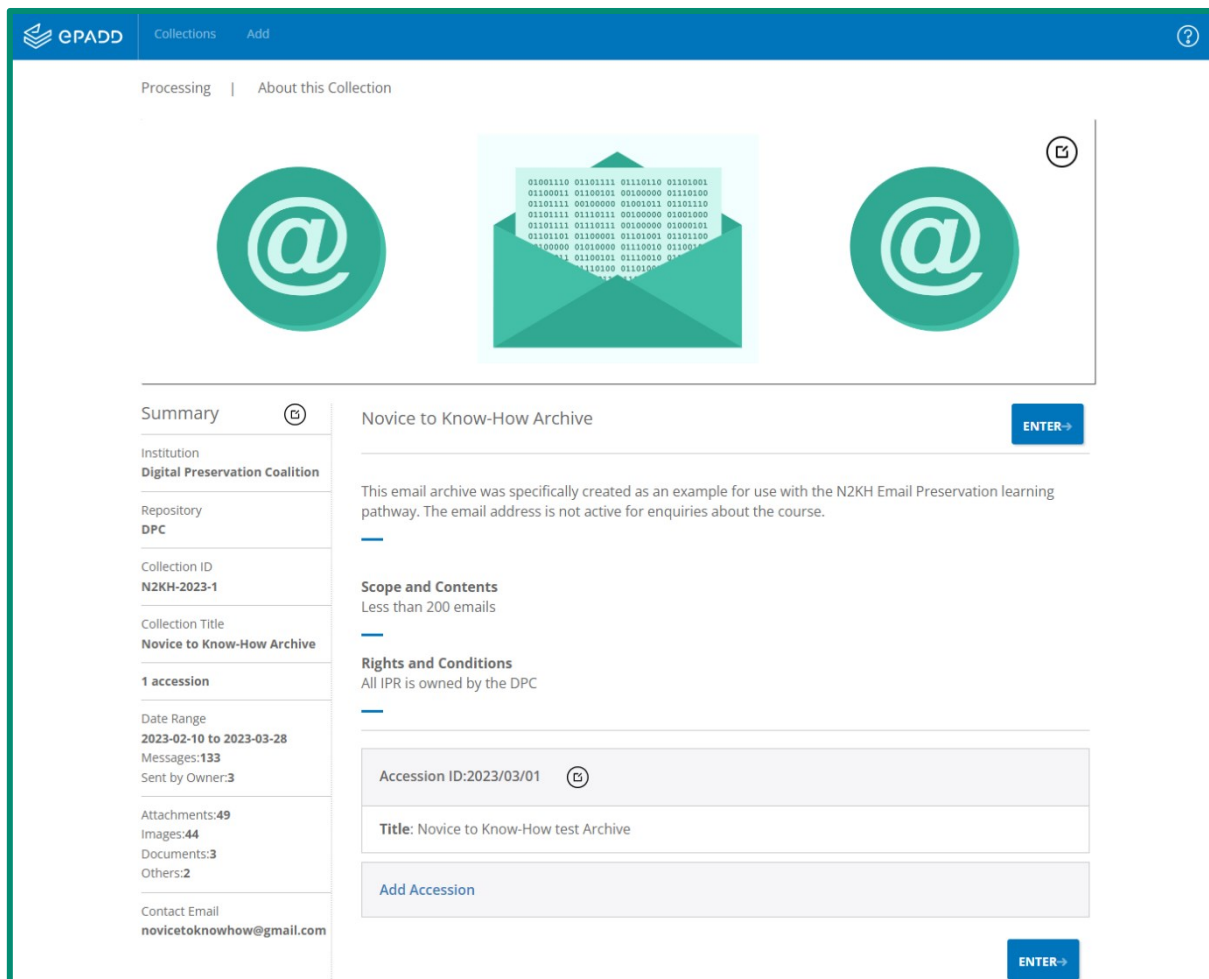
5) Click on the relevant tab to switch between **Collection**, **EAD**, and **PREMIS** metadata fields.



6) Click the **SAVE** button to save the entered metadata.

On the **About this Collection** page, you can also add a banner image for the collection that will be displayed alongside the collection description in the access modules by clicking on the  **Edit** icon on the gray banner space. The banner image should be at the aspect ratio 11:3.

It is also possible to add new accessions to the collection from this page using the **Add Accession link** and following the earlier instructions for importing content. A collection can also be deleted from ePADD using the **DELETE** button at the bottom left of the screen. There is no “undo” option in ePADD, so the choice to delete should be carefully considered.



The screenshot shows the ePADD interface. At the top, there's a blue header with the ePADD logo and navigation links for 'Collections' and 'Add'. Below the header, the page is titled 'Processing' with a sub-link 'About this Collection'. The main area features a large banner image with a green envelope icon containing binary code and two large '@' symbols. To the left of the banner is a sidebar with a 'Summary' section containing metadata: Institution (Digital Preservation Coalition), Repository (DPC), Collection ID (N2KH-2023-1), Collection Title (Novice to Know-How Archive), 1 accession, Date Range (2023-02-10 to 2023-03-28), Messages (133), Sent by Owner (3), Attachments (49), Images (44), Documents (3), Others (2), and Contact Email (novicetoknowhow@gmail.com). To the right of the banner, there's a 'Novice to Know-How Archive' section with an 'ENTER' button. Below this, there's a description: 'This email archive was specifically created as an example for use with the N2KH Email Preservation learning pathway. The email address is not active for enquiries about the course.' Further down, there are sections for 'Scope and Contents' (Less than 200 emails) and 'Rights and Conditions' (All IPR is owned by the DPC). At the bottom, there's an 'Add Accession' form with fields for 'Accession ID: 2023/03/01' and 'Title: Novice to Know-How test Archive', followed by an 'Add Accession' button and another 'ENTER' button.

After editing the metadata and collection banner image, you can return to the collection's **Browse Dashboard** by clicking on the **ENTER** button.

More on Labels

As mentioned earlier, you may wish to undertake further appraisal work in the Processing module, particularly if the work at the Appraisal module stage was undertaken by the original account holder. In the Processing module, all labels that were added earlier can be reviewed and edited.

At this stage you may also wish to identify emails that should only be transferred to the Delivery module. These will be emails that you want to provide access to only onsite in a reading room, with no information on them being made available through the online Discovery system. To do this the **Transfer to Delivery Only** label should be applied to relevant emails.

Export Options

Once you have completed final processing of the email archive, you will be ready to export versions for preservation, an Archival Information Package (AIP), and a version for access, a Dissemination Information Package (DIP). ePADD also offers export options for individual elements of the email archive, such as attachments and header information. At a minimum you should export a copy for preservation.

However, before exporting the archive, it is good practice to update the checksum for the archive to facilitate ongoing integrity checking. This can be done using the following steps:


- a) Click on the ⓘ **More information about this collection** icon.
- b) Click the **UPDATE CHECKSUM** button
- c) Click **YES** to confirm that the checksum should be updated.
- d) A **Success** dialog box should display. This can be closed using the **OK** button and the collection information box using the **CLOSED** button.

We will now look at the export options in detail. They are as follows:

1. **Export to Preservation** - you can select to export the archive in MBOX or EML formats. The EML option is only available if you have purchased an Emailchemy license and inserted the required license dongle. The exported MBOX file will be found at C:\Users\<username>\epadd-processing\epADD archive of <archive name>\data\exportableAssets\ProcessingNormalizedProcessed . This will be the version that will be included in an AIP.
2. **Export to next ePADD module** - this option will export the email archive ready for import into the ePADD Discovery and/or Delivery modules, ready for access provision. To complete this export, use the **BROWSE** button to open a file browser window and select where you wish to save the export file. ePADD will have autofilled the **Exported Archive Name** field, you can use the name as is or edit to your preferred name. Then click **EXPORT** to complete.
3. **Export Attachments** - if you wish to review attachments outside of ePADD, this option allows you to export them in bulk. You can select attachments by **TYPE** for common content types, or add specific format extensions for rarer formats. Then use the **BROWSE** button to select where the attachments should be saved, and click **EXPORT** to complete. *Note: this export option is not recommended for exporting attachments for preservation separate from their parent email as it does not automatically retain any intellectual links between the two.*
4. **Export headers (CSV)** - this option allows the export of header information from emails in a CSV format. This can be useful for anyone wishing to use the header information as data for analysis. To complete this process, select a location to save the file via the **BROWSE** button and then **EXPORT**.


5. **Download Messages** - this allows the bulk download of messages within one of three groups based on labels that have been assigned: **All Messages**, **Non-Restricted Messages**, and **Restricted Messages**. These can be downloaded as an MBOX file or individual EML files. As with Export to Preservation, EML is only available as an option if an Emailchemy license dongle is inserted. To complete this process, select the desired download format, select the type of messages to be downloaded, and click the **EXPORT** button. The messages will be saved to your standard **Downloads** folder.
6. **Export Entities** - this option allows the export of information about the named entities that have been identified within the email archive. You can download all of the named entity information, or entities within a particular group (e.g. Person or Place). To complete this process, select the chosen entity type(s) from the dropdown list and click the **EXPORT** button. The information will be saved as a CSV file in your **Downloads** folder.
7. **Export Correspondents** - as above for Export Entities but with the option to choose between **Confirmed Correspondents** and **Unconfirmed Correspondents**. Correspondent records can be confirmed when editing via the Correspondents list.
8. **Original text of all non-restricted messages as individual TXT files** - the purpose of this option is self-explanatory. As with the header CSV file, this may be useful for someone wishing to use the contents of the messages as data for analysis. Clicking the **EXPORT** button will save the messages to your **Downloads** folder.

Once your chosen exports are complete, you should save the archive one last time. You can then choose **Close** from the main menu to return to the **Browse Collections** page to begin work with another collection or close the browser window to exit ePADD.



Browse
Search
Authorities
Export
Save
Close

Processing | Export Archive


Novice to Know-How Archive

Institution: Digital Preservation Coalition
Collection ID: NZKH-2023-1

All Messages ⓘ

1

Export to Preservation
Emailchemy license: No license dongle present - Conversion for export to EML working in demo mode

SELECT
EXPORT

2

Export to next ePADD module
Specify Location

BROWSE

Exported Archive Name

ePADD archive of NoviceToKnowHow
EXPORT

3

Export Attachments
☒ Unrecognized by Apache Tika only
Type

SELECT
Other Extensions

Specify Location

BROWSE
EXPORT

4

Export headers (CSV)
Specify location

BROWSE
EXPORT

5

Download Messages
Emailchemy license:
No license dongle present
Conversion for downloading EML files working in demo mode
Download format

☐ MBOX ☐ EML

SELECT
EXPORT

6

Export entities

SELECT
EXPORT

7

Export correspondents

SELECT
EXPORT

8

Original text of all non-restricted messages as individual TXT files

EXPORT

Course 7: Discovery and Access

Module 7.1: Facilitating Discovery for Preserved Email

Introduction

Providing access to collections of preserved email starts with offering users routes into those collections, allowing them to discover content of interest. Due to the size of most email collections, traditional finding aids often do not provide the best option for discovery when used in isolation. Indeed, the keyword and full text searching offered by most email clients is far more powerful than what can be achieved through standard catalogs.

In this module, we will discuss how to approach the cataloging of email collections in a way that offers a balance between providing effective user discovery with the resources and time you have available.

A High-Level Approach to Discovery

As most email collections will likely contain thousands, if not tens or hundreds of thousands, of messages, attempting to catalog to item level is likely to be both time and resource intensive. With this in mind, it is recommended that, at least at first, you resist cataloging to this level, focusing instead on high-quality collection or series-level descriptions.

A good high-level description will likely offer enough information to allow a user to ascertain if the preserved emails will be of interest to them or not. Further discovery options can then be offered based on the access method used. If access to a complete mailbox will be provided for upload to an email client, the user will be able to utilize the in-built search facilities. If it is online access through the discovery module of an email tool or a repository access portal, there will be search and filtering options provided. These two methods for access will be discussed more in the next two modules.

A high-level description also offers the opportunity to place the email materials in the context of other related collection content, both digital and analogue, that the user may wish to consult.

Opportunities Offered by Tools

When preparing your collection or series-level descriptions, the information that can be extracted from an email account by email preservation tools can help with drafting many sections. These include:

- **Scope and Content:** Natural Language Processing (NLP) and Named Entity Recognition (NER) results will provide information on key organizations, individuals, and topics addressed within the emails.

- **Coverage Dates:** The tool will be able to identify the range of dates covered by the messages, including, potentially, times of high and low activity.
- **Extent:** Figures can be produced detailing how many emails are included as well as the number and types of attachments.

If you do have the capacity to consider cataloging the collection to a more granular level, there will also be opportunities for auto-populating fields using data extracted from the emails themselves, in particular from email headers.

The data produced by email preservation tools can also be used to create more innovative discovery aids. For example, data relating to sender and recipients has been used to create graphical maps of relationships between individuals within a corpus of emails. This can provide a quick visual guide to key parts of the collection that might be of interest.

Module 7.2: Providing Access

Introduction

The ultimate purpose of preserving emails should be providing access to the content at a future date. Users will likely wish to access preserved email for a broad range of reasons, some of which we highlighted at the beginning of this learning pathway including:

- Ensuring an organization's compliance with relevant regulations and legislation
- Offering emails as evidence in a legal case
- Providing information for a news article
- Contributing to research on an individual or topic

In this module we will look at some issues to consider when providing access and some potential access options you may wish to offer.

Issues to Consider

There are a number of issues to consider when providing access to preserved email. A key one relates to the access option that will best facilitate how users will want to use the emails.

At one end of the scale there are users who will want to access the email content in the same or a similar format to that used by the account holder. Here access options might include the following:

- Providing an MBOX file or collection of EML files that the user can import themselves into an email client, perhaps accompanied by instructions for this process.
- Email content preloaded in an email client for use. This may be the same client used by the account holder or one with broad compatibility you have chosen for access provision.
- Access to the emails through an emulator, perhaps of the original software used or of the entire computer environment.

An interesting example of the latter is offered by the work carried out at Emory University on the Salman Rushdie archive. An emulator was built to recreate the author's desktop environment through which researchers can gain access to born-digital elements of the collection, including email. [An instructional video relating to the archive](#) shows what this is like in practice.

At the other end of the scale will be users who are interested in analyzing the preserved emails as a dataset. For example, where they wish to map relationships between correspondents through the frequency of their contact and what is mentioned in the content of emails. For these users, provision of the preserved content in a format such as an MBOX file they can use with specialist software may be preferred.

Another important issue to consider is how much of the content will be made available to users, both in terms of size of the collection and the contents of the messages themselves. Decisions about the former may depend on the size of the collection and the number of messages included. For large collections it may simply not be possible to provide access to the complete collection at once due to the storage and processing requirements.

Decisions about both the proportion of the collection and the content made available may need to be made in relation to issues of sensitivity. You may wish to only provide access to parts of the collection where there are no sensitivity issues, or you may wish to only provide access to redacted versions of the emails.

Those decisions may also be affected by where access is to be provided: onsite or online. You may choose to provide access by one or both options and your choice will likely be influenced by resources and technology available, how you expect users will wish to work with the preserved emails, and if there are sensitivity issues. In the next two sections we will look at the benefits and drawbacks of providing onsite and online access.

Onsite Access

One of the main reasons organizations decide to only offer onsite access to preserved email is the greater level of control they can wield over the content, particularly where sensitivity issues exist. This can include:

- The ability to make case-by-case decisions as to how much of the collection and the content of emails will be provided to the user.
- The options for limiting the possibilities for replication and dissemination of the preserved emails to a third party. This can be implemented through actions such as deactivating external ports, e.g. for USB drives, so data cannot be copied and removed.

There can also be benefits gained by the user through onsite access. These include:

- The provision of software needed to access the preserved emails, that they might not have available to them. For example, a specific email client.

- Support from archives staff who are familiar with the collection and/or the systems and software required to access it.

The obvious disadvantage with onsite access is, of course, the limitations on who will be able to gain access. Only those in close geographic proximity or with the resources to travel will be able to gain access. You must, however, remember that providing any access is better than providing none and when you are just starting out or have limited resources, providing access onsite only is a perfectly acceptable goal.

Online Access

Online access provision to preserved email content is offered by a number of repository systems and email-specific tools such as ePADD. Online access provided through these systems can provide users with a range of useful functionality including the ability to search and filter emails, to browse collections by categories such as correspondent, and options to view graphical representations of the contents.

Another obvious advantage of providing access online is the ability to offer access to more, and possibly new groups of, users. This can also have the knock-on effect of increasing the profile of your organization.

Disadvantages include the need to potentially heavily redact or restrict access to collections where there are sensitivity concerns. There might also be limitations due to IPR issues. For example, you may only be able to gain a license from a depositor for onsite access only.

You will also need to consider the resources you have available to support online delivery of access. Such work will likely require collaboration with colleagues from IT, and will need ongoing support to ensure its sustainability. Without gaining resourcing to support online access, it may not be possible to provide it.

Module 7.3: The ePADD Discovery and Delivery Modules

Introduction

In this lesson, we will be looking at the functionality ePADD offers that supports discovery of, and access to, email archives. This functionality is offered through two separate modules, Discovery and Delivery, which we will examine in turn.

About the Discovery Module

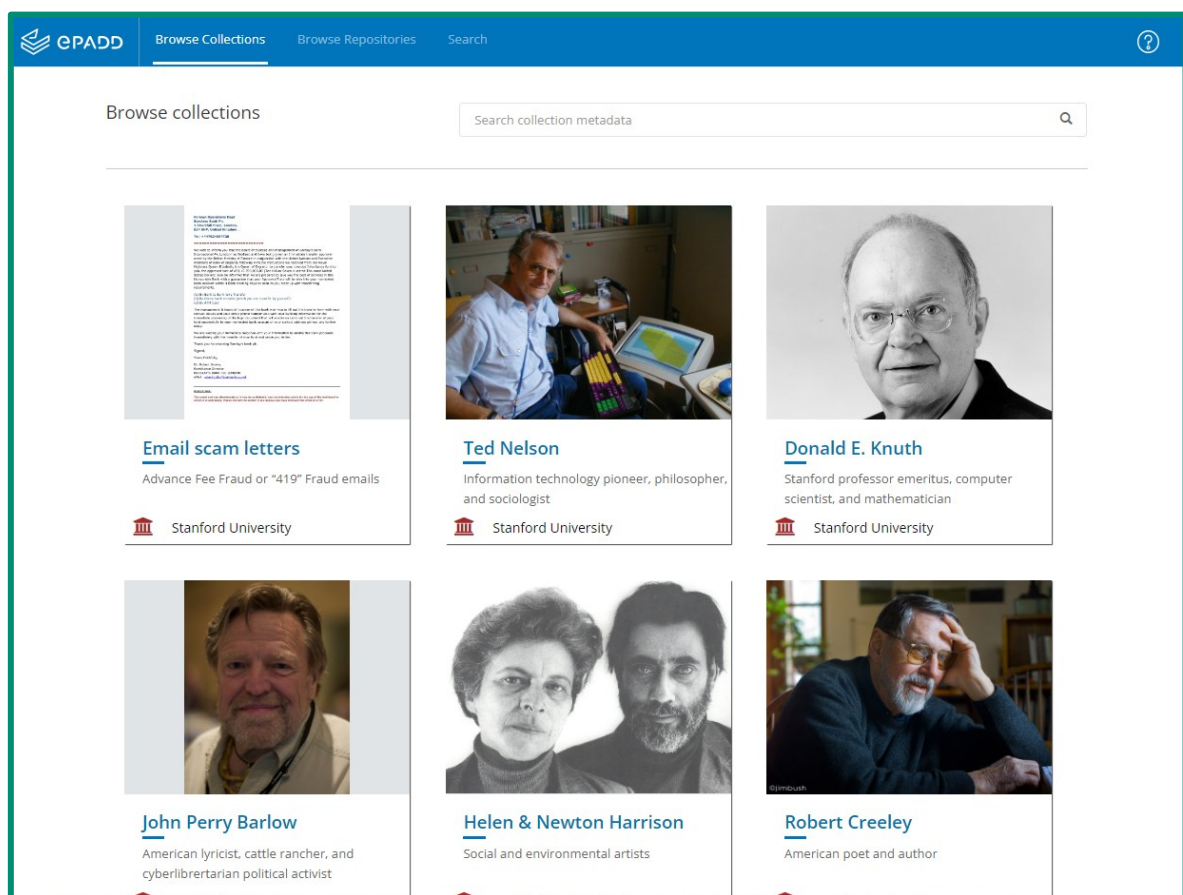
ePADD's Discovery module allows organizations to "remotely share a redacted view of email archives with users through a public web server discovery environment". It is designed to run under a web-server, allowing remote users to search redacted header info and extracted entities while limiting full-text access to the materials. This allows users to identify email

content of interest, whilst limiting risk of the release of sensitive information that may exist within an email archive.

Users accessing email archives through the ePADD Discovery module will be able to:

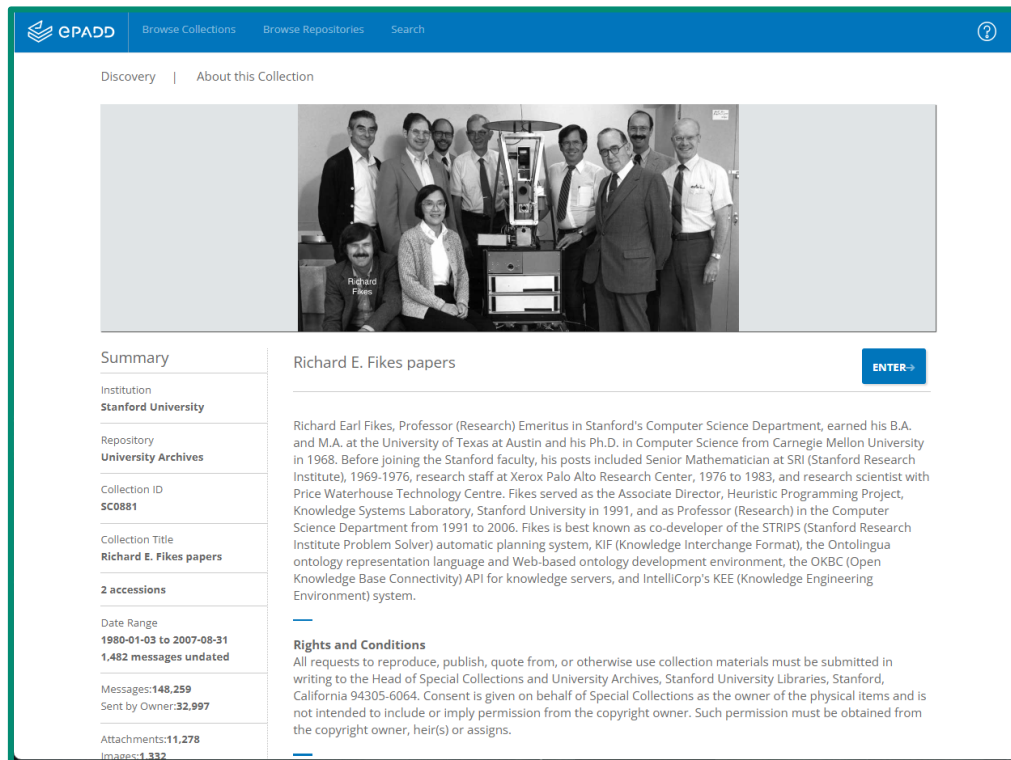
- Browse and search available email archive collections
- View collection-level descriptions
- Browse and search the contents of a collection by correspondent and named-entities
- View redacted versions of emails within the collection

Examples of pages from the Discovery module deployment for collections at Stanford University are included below.



The **Browse collections** page on Stanford University's ePADD Discovery portal:

<https://epadd.stanford.edu/epadd/collections>



Discovery | About this Collection

Summary

Institution: **Stanford University**

Repository: **University Archives**

Collection ID: **SC0881**

Collection Title: **Richard E. Fikes papers**

2 accessions

Date Range: **1980-01-03 to 2007-08-31**
1,482 messages undated

Messages: **148,259**
Sent by Owner: **32,997**

Attachments: **11,278**
Images: **1,332**

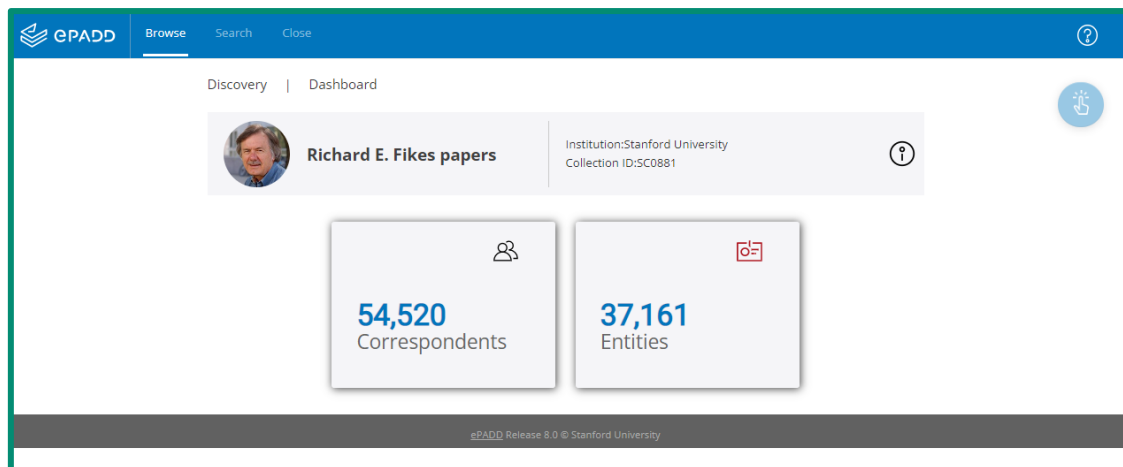
Richard E. Fikes papers [ENTER](#)

Richard Earl Fikes, Professor (Research) Emeritus in Stanford's Computer Science Department, earned his B.A. and M.A. at the University of Texas at Austin and his Ph.D. in Computer Science from Carnegie Mellon University in 1968. Before joining the Stanford faculty, his posts included Senior Mathematician at SRI (Stanford Research Institute), 1969-1976, research staff at Xerox Palo Alto Research Center, 1976 to 1983, and research scientist with Price Waterhouse Technology Centre. Fikes served as the Associate Director, Heuristic Programming Project, Knowledge Systems Laboratory, Stanford University in 1991, and as Professor (Research) in the Computer Science Department from 1991 to 2006. Fikes is best known as co-developer of the STRIPS (Stanford Research Institute Problem Solver) automatic planning system, KIF (Knowledge Interchange Format), the Ontolingua ontology representation language and Web-based ontology development environment, the OKBC (Open Knowledge Base Connectivity) API for knowledge servers, and IntelliCorp's KEE (Knowledge Engineering Environment) system.

Rights and Conditions


All requests to reproduce, publish, quote from, or otherwise use collection materials must be submitted in writing to the Head of Special Collections and University Archives, Stanford University Libraries, Stanford, California 94305-6064. Consent is given on behalf of Special Collections as the owner of the physical items and is not intended to include or imply permission from the copyright owner. Such permission must be obtained from the copyright owner, heir(s) or assigns.

The collection description for the Richard E. Fikes collection at Stanford University



EPADD Browse Search Close

Discovery | Dashboard

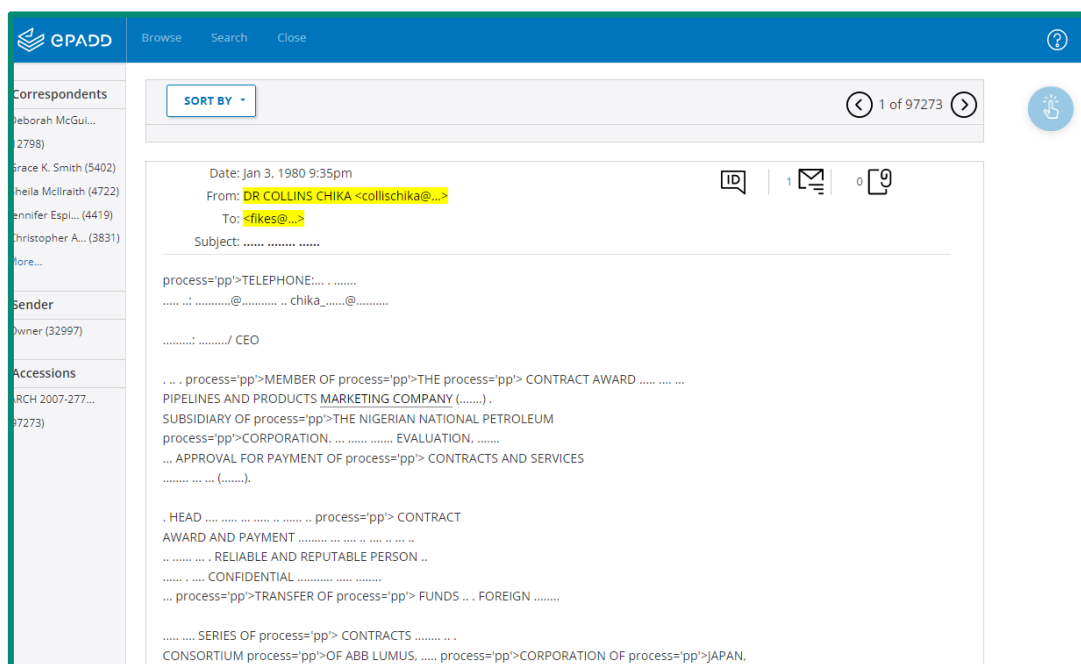
 **Richard E. Fikes papers** Institution: Stanford University Collection ID: SC0881

54,520 Correspondents

37,161 Entities

EPADD Release 8.0 © Stanford University

The **Browse Dashboard** for the Richard E. Fikes collection at Stanford University



An example of a redacted email from the Richard E. Fikes collection at Stanford University

Set-up of the Discovery Module

Set-up of the Discovery module will require mounting the tool under a web server and installation of the Java Runtime Environment on the same. Unless you have the web development skills and necessary access to complete this process yourself, it is recommended that you work with IT colleagues to install and set-up the Discovery module. Full instructions for this process are outside the scope of this learning pathway, but can be found in the [ePADD User Guide](#).

About the Delivery Module

The ePADD Delivery module “enables archival repositories to provide moderated full-text access to unrestricted email archives within a reading room environment on a local server.” Provision of full-text access onsite through the Discovery module will allow the archive organization to balance access provision with the risks of making potentially sensitive information available. Steps such as disabling USB ports, or limiting internet connections on the computers used for accessing email collections, can be taken to remove the risk of email content being copied and removed.

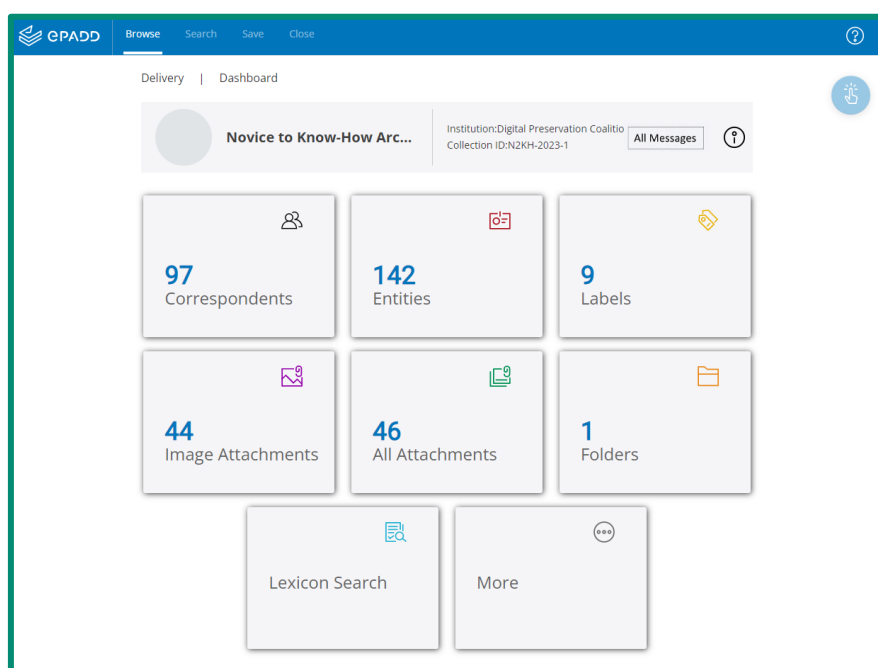
The Discovery module provides users with much of the same powerful browse and search functionality as is included in the Appraisal and Processing modules. This includes the ability to:

- Browse and search available email archive collections
- View collection-level descriptions

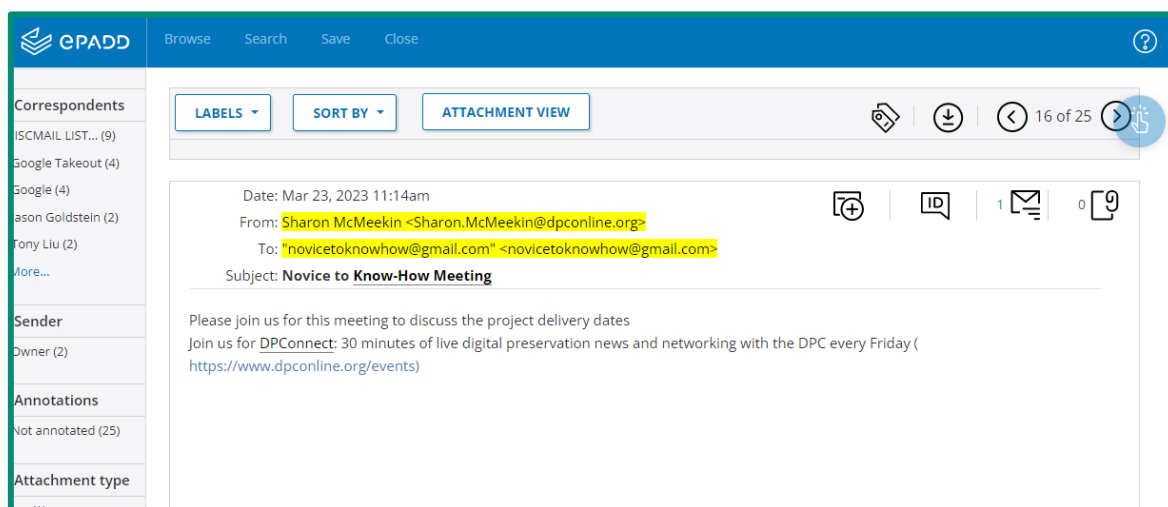
- Browse and search the contents of a collection by:
 - Correspondents
 - Named Entities
 - Labels
 - Image Attachments
 - All Attachments
 - Folders
 - Lexicon Search
- Export data on Correspondents and Named Entities
- Browse and view the full text of messages
- Add labels and annotations to track what elements of the archive have been examined
- Download attachments, individual messages, and groups of messages

Indeed, the main differences between the Delivery module and the Processing Module are that the Delivery module does not offer functionality for editing metadata, producing a report on the archive, and exporting content.

The **Browse collections** and **Collection Description** pages in Delivery look identical to those in the Processing module. The **Browse Dashboard** and **Browse Messages** pages, however, are quite different:



The Browse Dashboard of the Delivery module

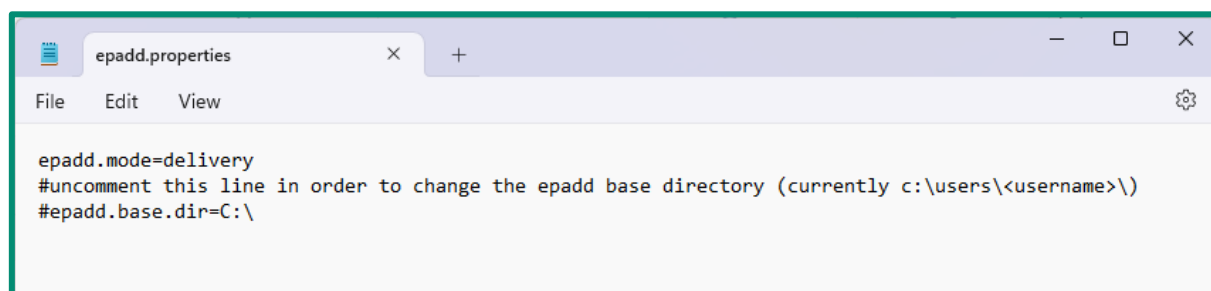


The Browse Messages page of the Delivery module

Set-up of the Delivery Module

Set-up of the Delivery Module on the relevant reading room computers will require the installation of ePADD and Java as described in the "Introduction to ePADD" lesson. Opening ePADD for the first time to the default Appraisal module will ensure the creation of all the necessary settings files.

You should then update the **epadd.properties** file in the **C:\Users\<username>** folder to set the Delivery module as the default (as shown below).



Finally, you must create a folder called **epadd-delivery** in the **C:\Users\<username>** folder and copy to the folder all email archives exported from the Processing module to which you will be providing access. These will have been saved to the location specified at the time of export and will be named in the format: **ePADD archive of <archive name>-Delivery**. It is important to copy the Delivery folder and not the Discovery folder as the latter will only contain redacted versions of the emails.

The Delivery module will now be ready for use, and the archives copied to the **epadd-delivery** folder should be visible on the Browse Collections screen when ePADD is next opened.

A few extra tips to keep in mind:

- It is wise to retain a copy of the Delivery versions of the archives in a location other than the **epadd-discovery** folder. As users can make changes such as adding labels, this will retain an “as processed” copy.
- If a previous user did make changes to the delivery version, you may wish to delete this version when they are finished and bring over a new copy of the “as processed” version of the email archive.
- If the “Reviewed” label was used as part of tracking the appraisal and processing of the email archive, you may wish to remove this label for all emails when providing access. This would allow the user to utilize this label to track their progress when working through the collection.

You may wish to provide users with a guide to using the ePADD Delivery module to help them understand the functionality it offers. All the content included in this learning pathway is available for reuse under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 License](#), so you can copy and adapt it for your own user guides.

Course 8: Email Preservation Case Studies

Module 8.1: UK Cabinet Office Case Study

Introduction

The Cabinet Office is a UK government department responsible for supporting the Prime Minister and Cabinet. It includes various units that support Cabinet committees and which coordinate the delivery of government objectives via other departments.

The Cabinet Office was the first UK government department to adopt a 'Capstone' type approach to email preservation, adapting the original concept used by the United States government to tackle email preservation within the UK regulatory context. This case study on their 'Capstone UK' approach is based on an interview with David Canning, Head of Digital Knowledge & Information Management, The Cabinet Office, UK.

Emails created and exchanged within this organizational context can provide important information and documentation of decisions by the UK government. At the same time, the identification, selection, and preparation of valuable emails for long-term preservation can pose unique challenges compared to other types of born-digital records, often when working with different approaches taken by different people and departments. As will be explained by David Canning in this case study, often "We are the trouble with email – it's that it doesn't really work well with human nature."

Background

The 'Capstone' approach, as explained in Module 4.2, bases appraisal and selection on the role or position of the account holders rather than the content of individual emails.

The Capstone approach was discussed in the 2017 [Better Information for Better Government Report](#), published by the Cabinet Office Digital Records and Information Management Team, working in collaboration with The National Archives and Government Digital Service. In the report, the authors give reasons why this 'keeping everything' retention approach may not be suitable for the UK system given significant differences in the regulatory context—particularly data protection and copyright law. Additionally, from a more practical and technological point of view, they expressed concerns about how to filter and sift through the large body of information that would be collected and retained under a Capstone approach.

While the 2017 Better Information for Better Government report concluded the Capstone approach might not be suitable for the UK, there were subsequent discussions and considerations by the Digital Knowledge and Information Team on how an adapted Capstone-like approach could be used for effective management and preservation of emails by addressing the two main areas of concern. The result of this work is what they call 'Capstone UK,' a Capstone-like approach that draws from the original concept with adaptations for their

governmental context and practical application for keeping and preserving emails at the organization.

In particular, the Capstone UK approach took into consideration

- General Data Protection Regulation (GDPR) and Data Protection Act compliance – How to ensure fair processing, balancing the rights of the data subject against the public interest in retaining important public records?
- Freedom of Information (FOI) compliance – Where is the most useful information stored for FOI requests to meet transparency obligations while not over-retaining?

There were also practical considerations about methods and processes – how to sift out and directly target the information they really wanted to retain and preserve for the public record. Implementing an effective sifting approach with filtering mechanisms is critical to help with

- processing emails in compliance with GDPR,
- identifying content that falls under data protection, and
- improving findability for FOI requests.

How they approached GDPR and Data Protection

GDPR created challenges to balancing the data subject's rights against the public interest in retaining important public records, but it also prompted the need and opportunity to develop a stated retention policy for a lot of digital materials, including email, and bring forth something productive to senior leaders to move forward.

Around the time GDPR came into force, little was being done with the piles of emails sitting and accumulating at the organization. Employees tended to pile up emails in their mailboxes because there was no data limit on the Gmail accounts in use. However, the Cabinet Office does not necessarily want to encourage employees to do too much tidying or deleting of emails because it leaves room for people to delete important information that should be filed and kept.

Other government departments took other approaches, such as automatically deleting emails after a specified period of time. This approach leaves room for the deletion of important information because it is based on the understanding that people will find and file the important emails before deletion occurs, but some people just don't do the filing and end up deleting records.

In both cases, the challenges relate more to people than the emails. The development of a retention policy is where the Digital Knowledge and Information Management team revisited the Capstone approach to create Capstone UK. The original Capstone concept lined up in several ways with their already developing records management policy. Whereas in the past they might ask individual users to make records management decisions, they agreed this would not work effectively with email as there would be individual users who did not make any records management decisions, and would implement them wrongly or inconsistently, all

leading to the problem of a growing 'digital heap' of emails that was hard to navigate through due to the inconsistencies. Where would they find the record in the email record?

The Capstone approach addresses this problem by no longer asking senior individuals, who would for the most part be operating through their mailbox, to make records management decisions for their emails. Instead of asking them to declare individual records, the entire mailbox of nominated senior staff and ministers is regarded as a record. All other staff are required to select individual emails that ought to be preserved and to store these, along with their corporate documents, in the shared drive. In this sense, the 'keep everything' aspect of Capstone was adopted for Capstone UK with a higher level (not individual) approach towards records management decisions. For less senior staff, the Cabinet Office operates a rigorous policy compliance and assurance regime that reduces the risk of important records not being preserved.

Capstone UK uses the grade of the individual, and their function to determine how far down the hierarchy the 'cap' should go. They retain everything at a ministerial, and permanent secretary level, and are more discretionary about what is retained for directors general and directors. This means only some emails are retained for the long-term record. The decision to keep is not a permanent decision made at the start with deposit—they conduct a first review, again after 7 years, and again at 20 years—so a decision to keep now is not a decision to deposit later in The National Archives for long-term preservation.

When an employee leaves the organization, there's a limitation period on employment tribunals requiring their information to be kept for at least 3 months in case an employee raises a complaint, or an employment tribunal. Conversations among the team about this 3 month period to retain employee emails raised questions about whether this was actually enough time to ensure that every single kind of investigation or issue would work its way out of the woodwork, so to speak, and ultimately they decided to set it at 12 months to ensure there was enough time to keep the information before conducting the first review. This period also acts as a further safety net to prevent the loss of important records in that it provides time for an ex-employee's work colleagues to raise a request to recover vital information.

How they used filtering and sifting methods and processes

Their current filtering process involves starting with a long list and then filtering it down further.

- They pull a report of all the mailboxes in the queue that are older than 12 months old,
- Pull a report from the HR Team of all those who are at a particular grade or level or higher (the 'cap'),
- From that designated 'cap' list identify the mailboxes to keep,
- Consult a list of mailboxes earmarked for a temporary hold, e.g. because of litigation or a public inquiry; and mark these for retention; then
- Arrive at a list of mailboxes proposed for deletion.

That proposed deletion list is then circulated around a number of subject matter, security, and data protection experts to identify any further mailboxes that should be put aside for delayed deletion. Most are typically approved for deletion, but this step is critical for identifying others that should be put aside, for example, because a FOI request or litigation has just been filed.

The retention policy understandably raised initial concerns about keeping personal data in emails, so it was important for the Digital Knowledge and Information Management team to establish and communicate how personal data would be removed, redacted or protected in statutory and professional terms. There were misunderstandings on personal data in retained emails—a misunderstanding that Cabinet Office would send everything in their mailbox to The National Archives unedited including personal data and personal emails—which is not the case.

First, they asked the senior leaders' private offices to, wherever possible, simply put a label in the mailbox called 'personal data' so that the email could be labelled and managed by them or another person (such as a private secretary or executive assistant), consulting and working with them. This allows them to flag and group the emails under a label the team can then delete and get rid of all in one group. It doesn't remove the human nature aspects of how much, little or inconsistently the label is applied but it does enable a way to identify personal data better and comply with regulations early on before emails are put in the archive. This is something they can work together on and encourage as a form of filtration after the creation of the emails, informed by someone acting contemporaneously in the mailbox.

To clarify any misperceptions of what kinds of personal data may be in these emails, David Canning explained that there is typically not a lot of personal data of a confidential nature in the email accounts of ministers. However, there will be more of this type in the mailboxes of senior staff. For example, it may be that "a Director General might in one breath send an email to the Prime Minister and another breath send an email to their kid's nursery. So we acknowledge that, and you're allowed to do that." At the same time, he adds that they do not and will not keep what has been given the personal data labels.

After the emails are put in the shared drive, the Digital Knowledge and Information team next takes on the major sifting and filtering for retention and disposal decisions before ingest, using reports and some automation. They take all that is put in the shared drive and put it through their filtration system to then decide retrospectively what the record is. In other words, the team does not ask "you, the business or you the individual user to declare records ever again" and instead makes the decisions on what is important or not important, falls under personal data and data protection or not, in a more systematic, consistent way in line with the records management and retention policy.

When reflecting on this during the interview, David Canning commented on how this was the best approach to retain everything above a ministerial, permanent secretary level:

"It was quite clearly obvious to us, because of the way that people work. Permanent secretaries and ministers all have shared mailboxes which are run by their private secretaries, and all of their work goes through that mailbox. So that's the record."

The team sees the benefit of retaining emails from directors and directors general, to record important information and records of their influence on decisions that may not go anywhere near a minister's mailbox, but are more discretionary about this process of determining what should be retained—it does not follow the same 'keep everything' approach as those at the ministerial and permanent secretary level.

The filtering and sifting for deputy directors and people just below is important because it is these grades of the Civil Service that often do much of the policy development work. They are the people doing the research, the thinking, and writing advice to ministers. The senior staff will become engaged in the backwards and forwards of discussions on policy development, usually via email, and may direct this in what is called 'thought leadership.' This can manifest itself in very long strings of emails in which, either through a direct instruction or a subtle suggestion, the direction of work is influenced. It is this aspect that mailbox retention aims to capture in the records, and which illustrates the unique dynamic of digital, revealing how people were talking to each other over a specific period of time.

The Capstone UK approach is not a perfect solution. There will be differences in how it will need to work for different departments because hierarchies and decision-makers will be slightly different, but in the Cabinet Office, it felt like a sufficient boundary for automatically retaining all mailboxes. The Cabinet Office do not, for example, automatically retain the mailboxes of deputy directors because they want the deputy directors to be the primary people driving compliance in their own teams, and lead by example. The DKIM team works to communicate the importance of good records management and puts them under pressure to save their important emails.

How they approached Freedom of Information (FOI) compliance

Compliance to Freedom of Information (FOI) legislation at the Cabinet Office involves a balance with data protection. They must ensure that the right information is stored and accessible to meet transparency obligations while also not over-retaining. With access, there is the need to comply with data protection while justifying retention on the basis of archiving in the public interest. There can be personal data in the permanent secretaries' and ministerial mailboxes where there is a strong argument for archiving in the public interest, while also having an argument for applying more discretion because of data protection.

One indicator to the Digital Knowledge and Information Management team that Capstone UK has been a successful approach has been from looking at the types of FOI requests received so far, like public inquiries where the targets of requests are almost always the higher-level: ministers, the permanent secretaries and directors. So far, there has not been any

request that has required the DKIM team to search for information in mailboxes specifically below these designated account owners.

Retained mailboxes are held in an archive for seven years from their closure, pending further review. At year seven the mailboxes will be further filtered to identify the information of historic importance that may eventually transfer to The National Archives. This further filtration may result in the deletion of emails within a mailbox, or of entire mailboxes, depending on the team's assessment of their value.

Reflections and discussion

Like the US Capstone approach, the Capstone UK approach by the Cabinet Office uses grades or levels to decide what emails should be deposited and kept, keeping everything above certain designated levels or grades. Unlike the US approach, not every deposited email must be retained for the long term. The Cabinet Office Digital Knowledge and Information Management team implements a mix of methods and processes to filter and sift the emails at various points—to help determine whether they should be disposed of or retained in compliance with GDPR and Data Protection legislation, accessible to meet transparency obligations and compliance with Freedom of Information legislation, and kept for long-term preservation at The National Archives.

In addition to the benefits of using this approach to help manage the emails, David Canning added that there is a business case for doing this. The Capstone UK approach has helped the system pay for itself by decommissioning accounts, and reducing the scale of unnecessary accounts and emails being added to the 'digital heap'. As he points out, the costs of using Google, like many other available modern cloud services, is less about data storage and more about paying for processing and paying for licensing.

"So every time we turn a mailbox from live into an archive, and every time we delete a mailbox we're saving money on licensing. When you're dealing with tens of thousands of users the savings can be significant, so doing nothing about it means that you're not just overstoring information, but you're being wasteful, paying licensing fees from taxpayers money unnecessarily."

Not only has it been easier to process and manage the emails through their Capstone UK approach and methods, but they have been able to document and demonstrate how they have saved a little over a million pounds in the past two years by getting rid of old accounts and eliminating the licensing liabilities.

By continuing to do this, with plans to review on a quarterly basis, they are able to see how the elimination of costs can create an opportunity to employ full-time archivist staff to develop and improve these processes even further.

Module 8.2: Trinity College Dublin Case Study

Introduction

The Library of Trinity College Dublin serves a number of functions, including legal deposit of British and Irish published works, and is home to Ireland's largest collection of medieval manuscripts and six million printed volumes. It is the largest library in Ireland with both analogue and digital collections.

This case study describes the approaches taken for a project to catalogue and preserve the first email archive donated to the Library. The email archive, comprised of 850,000 messages, is significant to the Library, not only as its first email archive but also as the largest born-digital collection in the Library's holdings.

The project began in 2021 and will conclude in 2024, and the case study is based on a March 2023 interview with Dáire Rooney (Project Manager and Archivist) and Brendan Power (Digital Preservation Librarian).

Background

The Library of Trinity College Dublin acquired their first email archive from an external donor organization who had approached them about taking their organizational archive. Originally the organization's emails were not considered but Library colleagues suggested it would provide a wealth of data about the activities of the donor organization.

This donation was accepted and acquired for its historical and cultural significance but was also significant as the first email archive collection at the Library.

The Library, which has been digitizing materials since 2007, holds one of the largest digitized collections in Ireland but it is only in the last few years that they have been approached about collections that are solely born-digital.

Due to the scale of the archive, the project was treated as a substantive pilot — to not only ensure the donated collection is preserved and accessible to researchers but also to help build capacity at the Library for accepting further born-digital collections and additional email archives they may acquire in the next few years.

Scope of project and this case study

The project was structured into two phases, with phase one focused on putting together a plan on how to tackle the processing and preservation of the material, starting with the identification of what was in the donated email archive. This would then help inform research and decisions about the approaches, tools and resources to be used in cataloguing and in the provision of future access.

Scoping was an important part of the first phase to get a sense of the size, scale and types of content in the donated email archive. A high-level summary of the archive was available prior to acquisition to give the project team an idea of the estimated size in gigabytes and numbers of files. Detailed scoping occurred after transfer and following the appointment of the dedicated project team, when the resource was available to further assess the material.

As the collection remains under embargo, information about the context of the emails, the donor, or the content of the emails could not be discussed by the project team during the case study interview. However, the discussion with the project team provided useful overviews to show how they

- Approached data protection and security issues
- Planned and implemented migration methods and processes
- Processed attachments and resolved issues with attachments

Data Protection and Security

The material in the email archive includes individual accounts for less than fifty employees, with a number of emails in those accounts containing potentially sensitive and personal data.

As part of the project plan, a governance group was created for the project team to work with and report to on potential issues relating to sensitive information and data protection. In addition to reviewing a Data Protection Impact Assessment (DPIA) conducted for the project, the College's Data Protection Officer was also part of this governance group, helping support work and ensure that what the project team was doing was acceptable from a data protection perspective.

Potential security issues were also raised and discussed early on in the project. Security was not just about securing the data but also focusing on things like potential viruses, especially when dealing with email collections. To address this, the project team shared storage and access workflows with the College's IT security team for review and they advised on how to effectively store data, limit access, and conduct virus checks to ensure viruses are not spread to the wider network.

While some of the specific steps taken in these workflows will be addressed in the next sections, at the most basic level, the project team made sure to:

- **Store data - Create multiple backups** of the original material in PST and MSG formats extracted from the original Microsoft environment in which they were created, with these backups stored in multiple locations, and on both external media and on an air-gapped network that is physically isolated from insecure networks.
- **Limit access - Create controls** to limit who could access the material in and across the different storage and locations. Access to the material was limited to specific individuals on the project and systems team. There were also location controls for access, meaning there could only be access via designated machines (approved hardware that is linked

to individual users) to diminish risks of unauthorized access to data through robust access controls.

- **Conduct virus checks – Use antivirus software** to detect viruses and files to extract or remove. These virus checks were conducted at different points in the workflows. This is especially important because the migration can sometimes expose viruses previously undetected during deposit and transfer, for example those may be in the form of zipped folder attachments being sent to individuals at the organization. Additionally, it was critical that these virus checks all occurred on a non-networked PC to ensure no viruses were spread to the wider network.

Methodology and Processes for Migration

The project team agreed it was essential to have the material extracted from the original environment to be able to create multiple copies of the content for preservation.

To accomplish this, migration was employed to extract the material and remove the external dependency on the email provider (in this case Microsoft) in order to create copies of the content that could be accessible and usable in different environments. The migration process also mitigated risks relating to data loss due to things like account holders being locked out of their accounts due to errors or data breaches.

It is also worth noting that the project team keep copies of the original files, with checksums, before any migration processes occurred. MD5 checksums were generated for all the deposited materials following receipt, as a means to ensure no changes or alterations were made to the original files and to check that the software used for migration did not alter the original files during the process.

Target Formats

To migrate the content to formats that would be interoperable with a much wider range of software beyond the Microsoft environment in which it was originally created, the team had to discuss what target formats would be best.

The main criteria for the target formats were that they are non-proprietary, open, and well established with a good track record in email preservation practice. They selected MBOX mailboxes and EML for the individual emails based on these criteria, noting how both are open text-based formats that are established and well supported by numerous email clients and email preservation tools—including ePADD, the tool they use.

Using ePADD

The project team chose ePADD for a few reasons, the main ones being it was user friendly and open source. Another benefit was how it worked well with the target formats, allowing the original folder structures in individual accounts to be presented in a similar way—to present

account mailboxes with the same inbox and original user-created folders used in the original environment.

They encountered some challenges in terms of uploading and ingesting. The project team explained how one of the challenges occurred at the beginning when the number of emails in each account mailbox could vary in size and scale. The mailbox of some accounts could have tens of thousands of emails that, even with a lot of RAM available, could still push processing limits and upload times—and on occasion crash.

Their solution to this problem was determining the point at which ePADD would be able to cope with a certain number of emails at a time (e.g. 10,000 emails at a time), and do the processing on a phased basis where accounts exceeding this number would have the emails brought together following the upload and ingest. This included incrementally ingesting multiple parts of the one large email account, having those parts merged within ePADD.

For example, for an account with 30,000 email messages, they would

- Ingest the number of folders that roughly made up 10,000 (e.g. Drafts, Sent Items), then
- Ingest the next 10,000 (e.g. Inbox), and then
- Ingest the final 10,000 (e.g. user-created folders).
- Once they are all in ePADD, they are merged and exported to the next processing module as a complete account.

Processing and resolving issues with attachments

As pointed out by the project team, when talking about the migration of formats for emails, migration usually refers to the message format rather than the attachments. The migration of the actual email formats, the PST or MSG formats, does not adequately ensure the accessibility of the content that is included as attachments to the emails.

They noted that ePADD places the attachments in a separate folder within the archive and uses a sequential numbering structure to align specific attachments with specific messages. In this way, the fidelity between the message and the attachment is maintained to ensure that there is an intellectual coherence between message and attachment.

The project team also noted difficulties with the range of attachments people have included with emails over the years. During their project, they encountered:

- Encoding issues where it could be a non-English language attachment, for instance, a file name with letters with accents that may not have been processed correctly.
- Files corrupted when they were initially attached to the e-mail, sometimes due to a failure to attach fully. For example, emails where someone sent one email with an attachment that is corrupted, followed by emails from recipients about how the file won't open, and then they attach it again—creating two of the same file but with one corrupted.

- Examples of corrupted files due to incorrect file extensions, such as attaching a PDF when it was a DOCX.
- Password protected files where they wouldn't have access to the password or, in some instances, people will have included the password in the body of the e-mail. This occurred with zipped folders that needed to be uncompressed.
- Issues where the files were nested too deeply within zipped folders, impeding extraction.
- Attachments which were executable files or software files, or system generated files or software that no longer exist.

The project team found there was not a way to solve a number of these issues with attachments within ePADD. So, in order to both maintain fidelity and solve the issues with these kinds of attachments, they continued to work in ePADD to maintain fidelity while also separately investigating and fixing attachments separately on their own using tools and techniques outside of ePADD.

One example of this would be a PDF file that, when looking at it within ePADD, seemed okay, but once downloaded would not open because it was not recognized as a PDF. The project team was able to identify that the PDF is really a DOC because they have tools for full format identification. Outside of ePADD, they can confirm this is a file identification error and document this in ePADD to note, 'if you want to open this, you need to change the extension to DOC,' indicating that a correct version of the attachment is available in a folder with attachments, that they would make on the access workstation in the reading room.

When resolving issues with attachments and preserving them independently in this corrective way, the project team has also made decisions allowing archival weeding or removal of files deemed unsuitable for long-term preservation. Sometimes this can include attachments like duplicate images of company logos, email signatures, letterheads, etc. For example, an email signature with a small Twitter image or Facebook image next to the signature shows as an attachment in ePADD, but they are independently able to separately preserve or weed image attachments like these to then allow them to produce more accurate statistics on the formats, and numbers of each format, of materials they are managing and preserving over the long term.

The information they gather in the process feeds into format watch and actions going forward to be able to act if any formats are in danger of becoming inaccessible.

The project team did, however, note that with duplicates, the retention decisions made by the organization, just as with other digital preservation decisions, should be context-dependent. One thing they noticed is that even though there may be a number of duplicate attachments, the information communicated through the multiples can sometimes serve different functions in different contexts. So even if there are five duplicates, for example, the five of them could each be fulfilling different functions—this is where they take into account the context of the

attachments in the folder, in their meaning or use, and creation when making any decisions about deletion of duplicates.

So, in short, there are complex issues arising with attachments that can lead to complex decisions about how to best resolve them: there is no “one-size-fits-all” approach. The decisions to keep duplicates have an impact on storage capacity and functional information, and other organizations should consider their contexts when making any deduplication decisions.

Reflections and discussion

The project ends in October 2024, but some positive outcomes have already been achieved.

As reiterated by the team during the interview, one of the initial aims was “being able to actually start processing emails.” They are now able to demonstrate how they have not only planned but implemented processes, to the point where they have over 200,000 emails processed for 16 of the individual accounts, which is a major milestone.

The progress on the project also shows there is a process model at the Library for acquiring and processing email and other kinds of born-digital material from donated collections, which can give stronger support for acquiring email and born-digital collections in the future. Or, as stated during the interview:

“from an institutional perspective, I think that having carried out a project like this is something we can also use as an advocacy tool for future acquisitions. So being able to have conversations with donors where initially perhaps the focus is on physical collections, but now we are able to ask a lot more questions about the digital component and think about their digital legacy as well.”

Within the context of the Library and Trinity College Dublin, the guidance and training shared through the project will demonstrate to colleagues how they can do this in the future by showing what is achievable in the context of current capacity, and the capacity needed in the future to develop in the areas of acquisition and preservation of born digital content. In this current phase two of the project, the team is working with colleagues within the wider department to provide training and guidance on processing born digital materials with the hope it will help ensure future planning after the project ends.

One of these milestones around guidance and training is access, working with reading room colleagues on guidance and training on ePADD and on a workflow for times when something goes wrong with a particular file.

There are also plans to work with curatorial colleagues in a similar way for guidance and training on workflows for acquiring and processing potential future collections, such as internal training through a webinar-like format and demos of ePADD to show how it works, how ingest is carried out, and what to look for when processing content.

Finally, the project team sees wider relevance by sharing how the principles have been applied, and the processes and workflows developed and implemented for wider application to other digital content at other organizations and institutions. One aim of the project was that the knowledge and experiences would not be siloed – that the knowledge from this project would have application beyond the specifics of the collection – so at the same time safeguarding this particular collection but also laying a foundation for future acquisitions to enhance the collection as a whole and acquire material in a much wider variety of forms.

Module 8.3: Sheffield City Council Archives Case Study

Introduction

The Sheffield City Archives, part of Sheffield City Council, collects, preserves and makes accessible materials relating to Sheffield and South Yorkshire. These materials include archive documents and records relating to individuals and various kinds of organizations—including local authorities such as Sheffield City Council and its predecessors.

This case study describes the email preservation approaches taken for the Sheffield Tree Archive project, a project undertaken by the Sheffield City Archives on behalf of the Sheffield City Council to catalogue all materials relating to its controversial management of street trees across Sheffield between 2015-2018. Not only was the project significant for its aims and motivations, but it was also the first time the Archives took on the collection, management, and access of email and born-digital materials. The case study is based on a March 2023 interview with Benjamin Longden, Archivist, who worked on the project along with Pete Evans, project lead and Archives and Local Studies Manager, and Clare Connolly, Archivist.

Background

Following the cutting down of trees across Sheffield between 2015-2018, elected members of the Sheffield City Council approached the Sheffield City Archives about undertaking a project to catalogue all material relating to the council's management of street trees from that period.

The work around the management of street trees proved to be a controversial issue for the city. In response, the Council's Cabinet decided to publish and make searchable as much of the materials as possible to help with rebuilding bridges with the community.

Project governance included a Project Board sponsored by an Executive Director, with senior representatives from the Council's Legal, Information Management, and Policy services. There was also an Operational Group, with representatives from Legal Services, Information Management and the departments responsible for managing street trees, from whom most of the information was to come from.

The project taking on the collection and management of the materials was essential to ensure the information and records contained in the digital content were preserved and accessible to the public. For the Archives, the project also served as a pilot for how they could assess and approach the preservation of emails and related born-digital content.

Scoping

Scoping was an important starting point for the project to try to get a sense of the size, scale and types of content that would be collected—from across the council and a range of stakeholders, including Amey, South Yorkshire Police, the Police and Crime Commissioner,

Forestry Commission, and the Local Government Ombudsman—to create a comprehensive and accessible collection.

However, scoping proved challenging for a few reasons:

- It was difficult to estimate how much material would come in as the materials had not yet been collected.
- The archives team was also responsible for collecting the material from the identified stakeholders, putting out appeals to various departments for them to deposit material that they could then appraise and process for the archive collection.
- The review and processing of materials occurred on an ongoing, continuous basis following deposit in order to make them available to the public as soon as possible.

For these reasons, it was difficult, if not impossible, for the project to have an accurate high-level estimate of the size of the collection in gigabytes and the numbers of files received, before undertaking transfer and preservation processes.

They had initially thought the project would receive and process the materials within 6-7 months. However, they received 10,000s of emails and attachments, and the project faced an unanticipated hurdle of having to plan, process, and make available a large amount of material in a short window of time. As explained by Benjamin Longden in the interview, it wasn't until the project progressed that the actual size and scale became clearer and, in light of what was discovered, necessitated scoping at various points before and after materials were processed.

Collecting and appraising the materials

The project began with collecting the materials from the identified stakeholders.

To do this, the project and archives team undertook presentations to Departmental Leadership Teams who then nominated key individuals to act as contacts between departments and the Archives. Some departments transferred relevant emails to a shared drive on the council network. For those departments that were unsure on what to search for, the Archives team did provide a list of search terms they could use. However, for other departments the team relied on the knowledge those particular teams had on what material would be most relevant. Others allowed the Archives team (within clearly laid out parameters and guidelines) to look through particular email accounts using eDiscovery software.

Before they transferred material, they carried out appraisal as they found a lot of duplication and out-of-scope material being deposited. This included the searching and selection of particular email accounts as well as looking through the deposited material in the shared drive.

In both instances, the selected material was next transferred to a *Master Copy* folder on a secured archives part of the network for designated staff to access and undertake the next steps.

All these emails were kept exactly as they were when transferred to the *Master Copy* folder. The files were kept in their original Microsoft formats with the original attachments included.

Receiving and converting the materials

Following the appraisal and transfer of the material to the *Master Copy* folder, the team then copied the emails, plus any attachments, into a created Catalogue folder, adding number references to the titles so that they were in date order. They converted the copied email and attachment files Catalogue folder into the PDF format. The reasons for selecting PDF as the target format included its interoperability to allow the content to be opened and accessed with a range of software.

Perhaps more importantly, given the nature of the materials, the PDF format allowed them to redact sensitive or personal data easily. The Council wanted to make redacted copies of emails available to the public, and redaction was not possible if keeping them in their original Microsoft formats. There are existing tools and techniques for redaction in PDF that could ensure effective redaction in line with the UK General Data Protection Regulation and the Data Protection Act 2018, the Freedom of Information Act 2000 and the Environmental Information Regulations 2004.

Practice Note:

The decision to use PDF was also shaped by project hurdles relating to the selection and procurement of a digital preservation software service. The team had initially identified Preservica as a digital preservation software service they were interested in to help store and optimize the findability and access of the emails and attachments. However, due to internal technical issues, the Archives team decided to adjust plans and work with what was available, which included the decision to convert and redact utilizing the PDF format.

Later on in the project, conversations with Preservica revealed that the software did not have the same capacity to redact in the way they required. Content would have to be redacted before adding the files to the software. Therefore their decisions to redact using PDF ultimately worked out for the best because redaction would have been required even if the software service was used.

The next stage was to combine groups of the files into one or more combined PDF files to prepare for the next stage of redaction.

Reviewing and redacting content

The combined PDF files were copied over to a redacted folder within the archives part of the network, where the files would next be marked up ready for redaction.

Using combined PDF files helped decrease the time and manual aspects of converting and processing 10,000s of individual emails, and it also helped with the redaction process. They found it easier to review and redact content in grouped emails—to work through combined

email messages presented in 100 pages of a PDF rather than open and work through 100 separate PDF files, especially when there can be a duplication of the same content within individual emails part of the same thread.

For the redaction process, they used Adobe Pro Plus 2020. Redaction included things like names, job titles, email and postal addresses, and phone numbers. Redactions were undertaken in line with FOI exemptions, such as personal data, commercial sensitivity and legal and professional privilege. A redaction protocol and framework had been agreed with the Project Board at the start of the project. This was published on the Council's website.

The project team and archives staff had existing capacity in terms of experience and knowledge working with sensitive and personal data, previously working with the paper archive of other high profile and controversial public events. Before any redactions were applied, at least one other member of staff checked through to make sure nothing had been missed while also trying to make as much material available to the public as possible. For example, they had to check that some of the material did not come under legal privilege or may be commercially sensitive, so they often had to check with the Freedom of Information and Legal Services Team.

Once it had all been checked, the redactions were applied and saved (removing metadata) to prepare the redacted PDFs for inclusion in the Calm catalogue.

Practice Note:

Redaction occurred earlier in the project than many might expect. They could work through the redaction process with already collected material that was publicly available while waiting for new material to come in. For instance, they went through the redaction process for documents that the Council had put on their own websites that had been taken offline at some point by IT. These documents had been available via the web, so they knew there had been some review of their content before they were made publicly available, but there were still some redaction elements that they could use to work through them as they were now historical documents with information that now could fall under UK regulations.

[Sheffield City Archives Redaction Guide](#)

Describing and structuring content

As noted by Benjamin Longden during the interview, a lot of thought and planning went into how to best describe and present the content of the archive to optimize discoverability and access:

"A massive feature of this project is that we and other council departments that passed on emails to us did not have access to individual email accounts but rather searched across individual email accounts using keyword phrases searches, and that was how we catalogued the material, while at the same time trying to recreate the folder structure within Outlook. So the collections are artificial collections in

the way they have been created out of the work of other departments and portfolios.”

They wanted the description and structure of the material in the catalogue to reflect, as best as possible, how it was in the original Microsoft environment. The catalogue has a series level description to represent the email inboxes and folders below with grouped emails, noting file size, file type, version, and arrangement within the corresponding folder.

They catalogued at a fairly high level, in groups of documents covering months or sometimes years, depending on the volume of the files, adjusting as necessary while working through the process. For example, there could be initial estimates of six months’ worth of emails based on what had been received so far. However, this might later need to be adjusted when the volume of emails corresponding to the period might actually end up being over thousands of pages and more unwieldy than expected. The descriptions were added to the Calm online catalogue to be findable and searchable to the public.

Practice Note:

There were early attempts to describe each individual email, but they soon realized the time and effort to do this was unsustainable. There were also occasions when describing at this individual email level where there was no clear subject heading in the email title and in this way, providing a useful catalogue description became more open to interpretation. This raised questions about how their interpretation of what should be included could be different from someone looking at it from another angle. Ultimately they decided that a ‘less is more’ approach was best, to give more time to dedicate to more pressing tasks and less on lower level descriptions that may not be helpful to—or interpreted in the same way by—those accessing through the catalogue.

Preparing and enabling access to content

The catalogue of materials is publicly available for users to browse and search via [Calm View online](#).

Before any access copies of the material are made accessible via Calm View, the project team works through a checklist in a spreadsheet to make sure all necessary steps are undertaken before creating and uploading the final PDF file for users to access. For example, one important step is to OCR the PDF documents to check if redactions were applied. If there is any personal or sensitive data revealed through the OCR, they remove it before creating the final access copy version. That final version is attached in Calm View.

Due to the volume and scope of the materials, they have uploaded and released them in a series of batches. The first batch of materials was made available in August 2022. At the time of writing this case study, there are currently 32 collections relating to the tree dispute in the Archive, bringing up 1,511 individual hits on their online catalogue.

Practice Note:

A lot of their decisions surrounding description and access to the content are, at their core, about managing public expectations. The project team is aware of both the strengths and weaknesses of how they approached the description and presentation of the digital materials for user findability and access.

Making the redacted content available to the public as soon as possible was the primary aim of the project, and in this way, the combined PDFs have allowed a more effective process for doing this. At the same time, one of the downsides of combining email messages occurs with email threads. Users may find a lot of repetition when scrolling through the PDFs of grouped emails which include email threads. Take, for example, a thread where Person 1 sends the first email to a group of people; Person 2 sends a reply-all message, and Person 3 replies to that message from Person 2. Depending on the context, the project team may just keep the third email for the access copy rather than keep all three separate emails with identical information. However, in cases where there may be a thread where someone comes in halfway through the chain—creating a kind of ‘parallel emails’ as phrased during the interview—they decide to keep that one while accepting there will be duplicate (but not identical) information presented by the email messages. This also is something they found to be the same when adding attachments as PDF, that users might lose that context because then it raises questions of “where does the email end and the attachments begin?”

Secondly, the higher-level archive description approach to cataloguing provides a structure to recreate the folder structure within the original Microsoft environment, offering users ‘artificial collections’ in the way they have been created out of the work of other departments and portfolios. In terms of searchability for users within the Calm online catalogue, there are limits to what they can do with this approach. This is often the case with traditional archival descriptions, in which users rely on the higher level descriptions to get a sense of the overall collection, but may be more apparent in these artificial collections of grouped email messages, where users must open and search within the files to see if the content in the grouped emails contains what they are looking for:

“I would imagine from a user's point of view there could be a bit of confusion too with search. You could search for a particular small residential street, for example, which you may not find until you've clicked in the PDF. You can imagine that they might look at the file and may feel overwhelmed. Or a user may not be able to find the street through the Calm search and is then not sure how to find it.”

Underscoring all of this is the ongoing debate about what is best to include—or not—when cataloguing everything in a collection. This is perhaps more visible when working with email messages and attachments. In theory, it would be best to have copies of individual email messages with item-level descriptions but in practice, this is not feasible and can become overwhelming when dealing with 10,000s of individual emails, attachments, and other born-

digital materials. The amount of labor to do this was realized early on, and given the unexpectedly large amount of materials received and the time-critical nature of the project, they chose the most practical and feasible approach to make the content available to the public as quickly as possible.

Reflections and discussion

Among the number of reflections and lessons from the project shared during the case study interview, there were a couple that stand out in terms of practical advice for others taking on similar projects:

Do as much scoping as possible early on

Scoping was an important starting point for this project but also challenging given the nature of the collection and the pressures to make the content accessible to the public as soon as possible. This is something that Benjamin Longden recognizes that “had it not been for the fact that elected members had publicly stated they wanted an archive that could be searched by the public into all matters regarding the tree dispute, I know we would not have taken this approach.”

While they could present an initial estimate for how long the project would take, it was only possible to accurately scope the actual size and scope of the project once they collected the materials from the various departments.

What was anticipated as a 6-7 month project grew and grew with the 10,000s of emails and attachments deposited and collected. In hindsight, he wishes there was more time to scope while also realizing it may not be possible as was in their case. For those who do have the ability or room to scope early on, he strongly recommends doing so to get a sense of the scale before you start.

Start small (and always save!)

There will always be a learning curve when taking on a new project. Despite the surprises they encountered with the size and scale of materials collected in the project, they found it was best for them to ‘start small and build up’ rather than ‘going right into it’ to implement all processes to all the materials together at the same time.

The next piece of advice offered is to begin implementing different approaches or processes on a smaller scale first, to allow opportunities to “see the pitfalls and then start to figure out and know what to do.” This is especially important to keep in mind when the task is bigger than anticipated. It is a ‘feat of endurance’ to do something on a large scale as they did, starting small first, allowed them to reach the point now where they can share thousands of documents and files with members of the public:

“It has been a very big job. There have been highs and lows, dead ends, false starts, and many, many technical glitches! That said, I have learned so much from this

project, and the public will certainly get a very detailed insight into the workings of the council, so something to take pride in, I think!"

When taking on those highs and lows, dead ends, false starts and technical glitches, another piece of advice is to make sure to always 'save, save, save.' This is especially important when falling into a working rhythm:

"As you get going, and you did more and more, you start to get a rhythm going. You get into a rhythm of planning the processes, and then applying them and finding out where you've got the problems—this is where to make sure to save, save, save. That's one of my biggest pieces of advice to carry along. Keep saving, because we had a lot of glitches along the way and you may as well."

Identifying opportunities and thinking ahead

One particularly interesting thing about this project is how it started as a time-specific project with an expected timeframe of months but with no rigid endpoint.

"This has been an unusual project. Normally it's like you've got a year to do it. You've got this amount of time, and then things change, and more stuff comes in. But normally, at least you have something to work towards. But this one's just grown and grown, really, and there's never been an endpoint."

The challenges faced also created opportunities to communicate the need and value of undertaking such a large project, allowing them to make a case for more resources and support to extend the project beyond the initial estimated timeline. Not only was Benjamin's Manager able to become more involved as the scale of the project grew, but they were also able to make the case to have a second, full-time person join the project in 2021.

The open-endedness has had benefits, especially in planning, implementing, and ensuring the proper procedures have taken place to provide as much access to content as possible to the public while also adhering to UK regulatory compliance and good practice. This is critical to show:

"With the redaction, we have curated and presented in a way that isn't what it originally was, and that does change the context. But we are able to show how we've done the proper procedures, so while it may not be ideal to users, it is the best we can make it available, and we still have the original to hopefully help build that level of trust. This is a huge question going forward particularly with this project because there was so much distrust towards the Council. So we've got to try our best to show we're doing everything properly, and that's another reason we want to make it all available as soon as possible to try and build those bridges again."

In this way, the project team also points to the necessity of setting an eventual endpoint getting the project completed as soon as possible— to then move onto other opportunities which may include preserving and making accessible other potential collections of emails and born-digital materials.

One opportunity they may revisit is procuring a digital preservation software service, particularly the one they had initially hoped to use at the start of the project to optimize findability and searchability through its search engine capabilities. Now that they have a process for redaction, it would be possible to bring in the already redacted files into the system, knowing they are compliant with regulations and can be made accessible.