CoSA PREPARE: PREPARING ARCHIVES FOR RECORDS IN EMAIL

SERI STATE ELECTRONIC RECORDS INITIATIVE
Council of State Archivists

Be Prepared

Managing and Preserving Email in State and Territorial Governments

Email Archives: Building Community and Capacity (EABCC) Grant, University of Illinois & Mellon Foundation

CoSA
Council of State Archivists

MAR
2023

CONTENTS

CoSA Council of State Archivists

The Council of State Archivists (CoSA) is a nonprofit membership organization of the state and territorial government archives in the fifty states, five territories, and District of Columbia. Through collaborative research, education, and advocacy, CoSA provides leadership that strengthens and supports state and territorial archives in their work to preserve and provide access to government records. CoSA facilitates networking, information sharing, and project collaboration among its member organizations to help state and territorial government archives with their responsibilities for protecting the rights and historical documents of the American people.

www.statearchivists.org

Support for this publication was provided by the University of Illinois/Mellon Foundation for PREPARE: Preparing Archives for Records in Email, a program of capacity-building services for email management and preservation that CoSA developed and implemented for state and territorial archives as part of the State Electronic Records Initiative (SERI).

SERI STATE ELECTRONIC RECORDS INITIATIVE

The State Electronic Records Initiative (SERI) was established in 2011 to focus on improving management, preservation, and access to state and territorial government electronic records in all 56 states, territories, and the District of Columbia. SERI increases capacity and capability for state electronic records management and provides a robust community of practice for archival staff preserving electronic records. SERI is committed to supporting the communities of practice, documents, guidance, and other resources to help state and territorial archives achieve their missions and increase equitable access to state electronic records. SERI promotes knowledge of the digital lifecycle and requirements for the preservation of state electronic records. SERI advocates for increased awareness and preservation of electronic records. SERI is the flagship program of CoSA.

INTRODUCTION

The Council of State Archivists (CoSA) received a two-year subgrant through the University of Illinois' Email Archives: Building Capacity and Community program,[1] funded by the Mellon Foundation to develop and deliver a variety of capacity-building activities for email preservation and access. PREPARE (Preparing Archives for Records in Email) was envisioned to deliver ongoing learning, information exchange, and collaboration among state and territorial archives with the broader archival community. PREPARE builds upon digital preservation research and analysis conducted by the archival community over the past decade, including past work by the Council of State Archivists.

PREPARE was designed in three phases: the project began with Phase 1, a needs-assessment survey to state and territorial archives regarding email preservation capabilities in state/territorial government. Phase 2 consisted of the organization of a focus group of expert digital archivists and information technologists from across the US to test email tools and report on findings, while Phase 3 focused on synthesizing the knowledge gained and turning it into direct assistance to states and territories.

The Council of State Archivists would like to thank:

- The University of Illinois and Mellon Foundation for their Email Archiving: Building Capacity and Community grant program.
- All the members of the advisory group that aided PREPARE in developing its needs-assessment survey.
- The members of the focus group of digital archivists who contributed to Phase 2 of the project.
- Volunteers from across the US who contribute to the State Electronic Records Initiative.
- Nicholas Connizzo, SERI Consultant and principal author of this report.

For Phase 1, in the spring of 2021, each state and territorial archives was provided with a detailed survey about email preservation maturity and capabilities, obstacles to email preservation, and capacity-building goals for the near future. For full results, and an analysis of the survey data, please see the PREPARE Phase 1: Needs Assessment Report, but some highlights are shared here.

State and territorial archives surveyed identified the biggest obstacles to efficient and effective management and preservation of email as:

- Quantity of email records (66% of states listed this as a challenge and 19% identified it as the biggest obstacle faced)
- Insufficient statewide policy adoption (63% and 16%)
- Lack of technology (63% and 13%)
- Lack of stakeholder collaboration (42% and 9%)

High-priority goals identified in the PREPARE needs-assessment survey were focused primarily around management and preservation activities that would ensure permanently-valuable email makes it to the archives (or designated repository) and can be retained there as long as necessary. The most common concerns were governance and policy adoption, transfer of email to the state/territorial archives, and implementing stability in email preservation workflows. Access was, comparatively, a distant concern and with good reason: few state and territorial archives possess email collections, and those that do have conducted relatively few processing actions. Hence, the need for PREPARE.

The results of the survey guided the next phase of the project, tools testing, which began in Fall 2021 and ran through spring 2022. In this phase, a group of a dozen digital archivists and information technologists from around the US (representing 8 states) tested email processing and preservation software tools in a closed environment with a large sample set of test data. For a full discussion of the focus group membership, tools used, and testing process, please see: PREPARE Phase 2: Tools Testing.

PREPARE created a digital preservation tool testbed with Linux- and Windows-based virtual machines (VMs), which were then distributed to the individual members of the testing team for installation and configuration on their local devices. PREPARE's goal was to create a stable, consistent environment that could be used in a wide variety of state computing environments and provide consistent experiences for the users. Along with these VMs came a sample set of email provided by the Utah State Archives, as well as a pre-installed and pre-configured set of email preservation and processing tools, primarily ePADD, libratom, and DArcMail, with some other tools to convert and manipulate records (such as Mozilla Thunderbird to convert email from one format to another).

Testers were given three months to work with the sample data and answer some survey questions about the tools' capabilities, efficacy, and most importantly, suitability for inclusion in government archives' email preservation workflows. Basic processing workflows were provided, but testers were encouraged to use their own workflows and email records, if desired, in order to provide as close as possible a real-world scenario.

The testing group found significant utility gained from each of the tools surveyed, but generally struggled with some technical expertise required to properly operate and troubleshoot the test software, and found difficulty in fitting these tools into existing workflows. Specifically, without predefined workflows (that include both inputs and outputs), testers found difficulty knowing *what to use the tools for*, rather than how to use the tools. Thus, the resounding message from the testing group was a desire to learn or collaboratively develop a standard-of-practice around email preservation.

Phase 3 of PREPARE involved taking the lessons learned from Phase 1 and 2 and applying them in real-world situations with volunteer archives to build capacities around email preservation. During the call for direct assistance, a number of states and territories expressed interest in the project, and many of those projects are ongoing at the time of this guide's writing. While this guide will not report on the specific progress of those projects, please check the State Electronic Records Initiative site in the future to see more resources that have come out of these direct assistance collaborations.

Instead, *Be Prepared* is intended to be the final piece of indirect assistance that will be provided to state and territorial archives, information technology, and records-producing agencies, and the digital recordkeeping community at large. It is intended to serve as a broadly-comprehensive guide to managing email records in state and territorial government, and combines and contextualizes much of the work that came before. *Be Prepared* is designed to be modular, and provides an overview of the common current state of email management and preservation programs, along with outstanding issues or concerns and recommendations to build sustainable programs throughout government.

A Semi-Technical Overview of Email

To be prepared to manage and preserve email, administrators, archivists, and records managers in state and territorial governments must first build into their enterprise governance frameworks a thorough understanding of what email is, what its management requirements are, and how to assign roles that properly distribute responsibilities to ensure that the email lifecycle continues to function smoothly.

> *The challenge of email preservation in government is primarily one of governance, rather than technology, but it is certainly exacerbated by factors of scale and the development of email format, transmission, and application evolution over time.*

The challenge of email preservation in government is primarily one of governance, rather than technology, but it is certainly exacerbated by factors of scale and the development of email format, transmission, and application evolution over time. This guide will provide a simple technical overview of email generally, including the most important factors to its preservation, but should by no means be considered to be exhaustive in its treatment of the technical aspects of email preservation.

For a more thorough technical overview, with a discussion of some of the more challenging aspects about email preservation, PREPARE recommends *The Future of Email Archives: A Report from the Task Force on Technical Approaches for Email Archives.*[2] Additionally, Preserving Email (2nd Edition) by Christopher J. Prom is another valuable resource for developing a broad understanding of email and its preservation challenges.

Overview: Anatomy of Electronic Messages

Ultimately, email is not a complex electronic record. It is mostly self-contained (though it can include content embedded from other sources, as will be discussed later). It has a discrete trail of provenance; sender, recipient, and the time at which the message was sent and received are all core components of the electronic record. It has a robustly-defined syntax and set of specifications that were designed to allow compatibility across different platforms, software, and systems. Email messages are almost all text, with the MIME specification allowing the attachment of other content types as part of the transmission.

On the other hand, some of these aspects of email can prove to be drawbacks as well as advantages. Only several header fields are required, and the same simple syntax allows for creating custom fields, which has historically resulted in a wide proliferation of dozens or possibly hundreds of potential email header fields to consider as sources of essential information or context. Furthermore, this has rendered many format migration tools (that purport to convert Outlook MSG or PST files into open-source EML, XML, or MBOX files) unreliable due to the frequent inability to detect and convert these custom fields.

What is email?

Some users refer to email as a platform or a service, others use the term to mean a single message, while others still use it to refer to a set of messages (*"I'm going to look through my email."*).

A single email is a specific kind of electronic message. Specifically, it consists of two parts: an envelope (whose specifications are described in the Simple Mail Transfer Protocol standard, a.k.a. SMTP)[3] and a message (as defined by the Internet Message Format standard, a.k.a. IMF).[4] This definition disambiguates between email and other kinds of electronic messages, including text messages (which use a different standard, the Short Messages Service (SMS) or the Manufacturing Message Specification (MMS)), direct messages sent within a closed system (such as Microsoft Teams, Google Meet/Hangouts, Twitter, Facebook), digital forums (Discord) and chat protocols (IRC).

> *A single email would be trivially easy to preserve; but an entire mailbox, with many interconnected records and their individual attachments, is a different task altogether.*

Emails individually are typically not large, or complex records. They consist, in short, of these primary elements:

- Message Headers
- Message Body
- Attachments

Of course, the task of preservation is made more complex by the staggering quantity of email. A single email would be trivially easy to preserve; but an entire mailbox, with many interconnected records and their individual attachments, is a different task altogether. Successful governance and preservation of each of these data elements individually, as well as

preserving any contextual links within or between these data elements (or between similar elements of different records), is essential to the task of digital preservation of email in state and territorial governments.

Building Blocks: Internet Message Format and Multipurpose Internet Mail Extensions

The IMF specification originally defined two components of a message: the header and the body. The message body was initially assumed to be flat US-ASCII encoded text.[5] The header of the email is a predefined set of metadata fields appearing in a specific syntax. Headers are, essentially, embedded metadata fields within the email, and the specification defines what these are and what forms they must take. These headers are followed by the body of the email, which has some required characteristics of its own. *Simply put, if an electronic message does not conform to this standard, it is not an email.*

For a full set of field definitions, further reading of the IMF specification is recommended.[6] *Be Prepared* will not address every possible field, whether officially registered[7] or not. The prevailing reason for this is that there is potentially a very large quantity of potential email header fields, as individual email software can create a limitless number of header fields as part of the message, as long as the required headers are present and the message is structured properly.

Importantly, only some of these header fields contain crucial information about the context of the message.[8] The following header fields[9] are of particular interest to preservationists in government, as they represent the places where the most valuable information is retained:

- **<orig-date>** and **<from>:** Only two message header fields are required syntactically: an origination date (orig-date) and an author (from). From a preservation perspective, this metadata absolutely must be retained to maintain the digital provenance of the record. Thus, any migration, export, or conversion process that fails to preserve these headers will damage the value of the record irreparably.
- **message-id:** While this is not a field required by the specification, it is extremely common and the primary enabler of the "email thread" function in modern email applications (headers track information that indicates when a message is a response to another message in this field). Importantly, the message-id field should include a globally unique message identifier issued by the email server that is never to be repeated. For reasons of digital provenance, in contexts where emails should have contextual links to their originating server, this is an essential piece of data to preserve.
- **Informational fields:** Optional informational fields, including *subject, comments,* and *keywords,* intended for human-readable content. While these fields may not appear, if they do they are likely to contain valuable information.
- **bcc:** Blind Carbon Copy (bcc) requires some special attention. A recipient of an email containing this field will (most often) only know the recipient <to> and their own inclusion (bcc). However, the sender's version of the message will list all blind carbon copied recipient addresses. Therefore, if *bcc* fields are commonly used, preservationists should place additional emphasis on the retention of sent emails as they will be necessary to trace the provenance of all recipients of a particular message.
- **Custom headers:** Most often designated by a preceding X, custom email headers can be used for as many purposes as email server software designers can imagine (message tracking is a common example). Custom headers are not guaranteed to be recognized by the recipient email server, and are not guaranteed to be retained (e.g. they will be "stripped") when forwarding from the new server. Preservationists should research all custom headers (really, *all headers*) in use by email server applications to understand what kind of data is expected to be in each sent email. Custom headers can contain essential metadata, especially when it comes to digital provenance, including essential information about the sender or the message. For example, certain kinds of digital signature mechanisms can utilize custom headers to facilitate the signing or authenticating of the message.[10]

Of course, all email headers contain information which is valuable and *Be Prepared* recommends retaining as many as are practicable given existing technology and resources. Email headers are not always presented to the users of email applications, so can contain "hidden" information, mostly technical, that can be extremely valuable in government applications. Dates and times, specifically, of when emails were sent and received by email servers, forwards, and so on, can provide a detailed set of evidence related to administrative decision-making that is crucial to public recordkeeping.

Email Message Bodies

Originally, email message bodies consisted entirely of text. While the specification did not originally provide for anything other than ASCII text in the message body, subsequent revisions to the standard and, more importantly, extension of the IMF specification to include **Multipurpose Internet Mail Extensions (MIME),**[11] allowed for the inclusion of data types beyond text, including alternate types of text such as HTML and CSS, fonts, audio, video, and more. Furthermore, for email bodies that make extensive use of external links to load or insert content, that linked information could itself be meaningful in and of itself (example: signature blocks included as an image could contain significant identifying information about the sender). This is entirely separate from an attachment, which is a discrete file that is included along with the message and may be the entire message body or may be one of many.

Preservationists cannot necessarily ignore details such as fonts, text size, coloring, and any other graphical representations that may be present within email bodies, even if these are less valuable for collections of administrative records

over individual manuscripts. These could easily be significant (example: *"please see the highlighted section in the next paragraph"*), or they could be expressions of user preferences that are not likely to have much long-term value. There is no single answer about how to appropriately capture this kind of content.

Either way, it may be exceedingly difficult to make value judgments such as these across the entire enterprise of government—especially when some of these examples above involve records creators that are outside government, and therefore over whom those involved in email lifecycle management have little control.

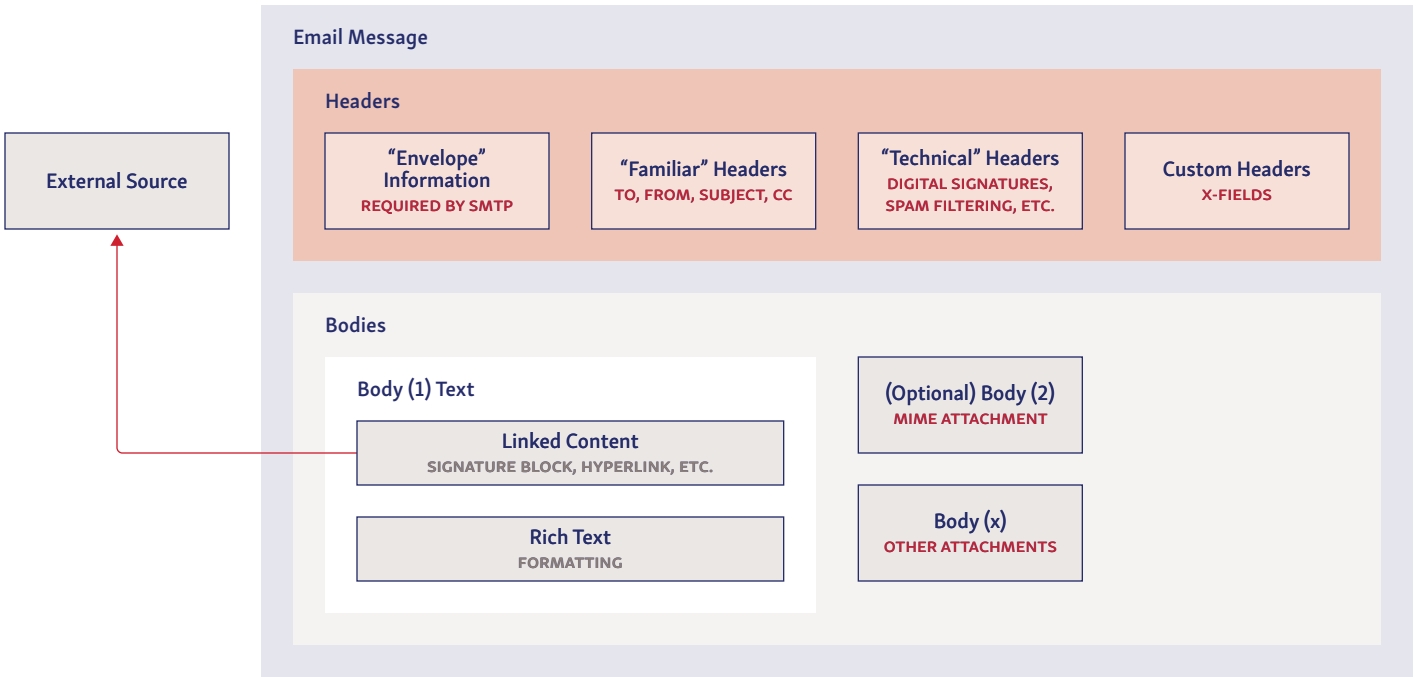
For these reasons, given the scale and scope of email in state government, Be Prepared recommends that state and territorial archives retain email in its original, unaltered form whenever possible. Migrations or normalizations should be made only in ways that retain all headers, body content, attachments, and any other contextualizing data whenever possible.

Email File Formats

Emails can be sent and received through SMTP by server applications as long as the data conforms to the IMF standard; but this standard does not specify how that data must be stored at rest, or how it is to be presented to the end-user. These are design decisions left to the email application. Because of this, email data has been stored in a variety of formats, some of which are:

- The **Electronic Mail Format (EML)** is an open format that was essentially originally specified in RFC 822 and was updated with the industry standard internet message format (IMF), which is the standard syntax for electronic mail transmission. EML has several advantages beyond its openness: it is supported by most email clients (and is likely to continue to be), it can be read outside of dedicated email clients (including a text editor native to essentially any operating system or platform), and it can be easily read by HTML viewers. EML is not much more than plain text that conforms to the specific syntax described in IMF. As a drawback, EML does not provide for the embedding of attachments (though EML messages will include in their message body data all MIME-related data including the filename, MIME-type, and so on). EML also does not embed external content (fonts, images, etc.) of any kind (though, again, links to such may be provided), so care will need to be taken to preserve links between the EML and those records.
- While **MBOX** can refer to a family of related file formats (MBOXO, MBOXRD, and others), its primary function is to store messages of a single folder (not necessarily a mailbox, so not multiple folders) in a single database file. MBOX files are useful since they collapse all messages into a single flat file, making it easier for applications to parse large swaths of data with few intermediate steps. MBOX is similar to EML in this respect, being somewhat akin to a "flat file" of text, with individual messages appearing in sequence. One limitation of MBOX is the historical lack of strict formatting syntax "standards", thus, each application that creates MBOX files will do so slightly differently.[12] This can cause compatibility issues in the future if care is

Anatomy of an Email Message



not taken to document what software is used to generate MBOX files, and which elements are captured in the file.

- **PST/OST/MSG (Microsoft)** are commonly-found file formats in state and territorial governments given the ubiquity of Microsoft Exchange, Outlook, and Office365 as their email platforms. These formats are openly documented[13], but present long-term preservation challenges of their own. PST is the most convenient and useful, as it is a self-contained archive format that can carry a great deal of information about messages, as well as the account itself. However, migrating data from these accounts is difficult, and Microsoft provides few official tools for migrating to open-source formats given the complexity of the data.
- **PDF/A** is a common destination format for migrations and conversions from a variety of file formats, and email is no different. PDF/A already supports embedded files which can allow for the inclusion of the original message, attachments, and potentially other linked content (such as embedding fonts or documents from external sources). Furthermore, email headers can be migrated to XMP metadata within the PDF file (or rendered as a page in the PDF) so that there is minimal loss of data, and to retain some of the original functionality of the record.
- **XML** schemas (like the one utilized by the Smithsonian Institution and others through their work to preserve email through the Collaborative Electronic Records Project) can be a useful format, since most of the data in an email is highly structured and lends itself well to conversion; an XML schema that includes all default email headers, body, and data about attachments such as filenames, size, and locations can facilitate preservation of email. XML is highly extensible that it can also be used to store contextual information about the account, as well as all messages in a single document.
- **Other formats:** Since email data is mostly structured text and MIME-type attachments, any format that can store and render this kind of data could be suitable for the retention or preservation of email. However, without significant analysis of what any format does better than those listed above, it is not recommended to migrate or store email in other formats.

Technical Challenges to Email Preservation

Many of the technical challenges to managing, preserving, and providing access to email records have been discussed above, but they are also presented here in summary format. *Be Prepared* hopes to address these challenges through specific recommendations, included in the next section. A summary of the technical elements to email brings us to the primary challenges of email preservation:

- Government officials use their email accounts to conduct public business in multiple contexts and across multiple functions, especially those involved in high-level administrative processes that are more likely to contain permanent or archival records (such as elected officials, agency heads, and senior advisers). Sometimes, information governed by different records schedules (and with differing requirements) is contained within a single email record. **Simply put, preservationists must assume that email is used by everyone, for everything.**
- Most email has very little organization, description, or indexing. The work of organizing, describing, and indexing email manually is time-consuming and unlikely to be performed by high-ranking staff whose emails are likely to require retention and preservation. **Tools available to accomplish this work at scale are limited or proprietary.**
- Beyond the requirements of the IMF, SMTP, and MIME specifications, email servers create, store, and send email in unique ways, primarily through customized email headers. **No two email servers operate identically, including different versions of the same software, and messages sent and received by different programs can look quite different.** This can be an asset (allowing deeper understanding of technologies used) or it can be a drawback (damaging to interoperability and especially to format migrations which can be tripped up by custom header metadata).
- Transfer or export from enterprise email systems can be resource-intensive, and can degrade the availability or reliability of other processes (especially FOIA/public records requests).
- The most commonly used email server systems in state and territorial governments are Microsoft Exchange and Gmail. These systems are designed with service-based licenses in mind, with data classification and retention as secondary priorities. **Migrating data out of these systems can be cumbersome, and can lead to data loss without detailed migration and export plans.** Microsoft systems, notably, have few native methods to migrate data in bulk to non-proprietary formats; Gmail will allow the export of data into MBOX format.
- **Email format converters are notoriously unstable, and each may produce different outputs even with the same input.** Some will strip most email headers, others strip custom headers, while others will strip any header not recognized. Careful attention must be paid to any software that purports to "convert" email from one format to another to ensure that no data is lost.
- Attachments stress storage space requirements and need to be preserved according to their unique characteristics (which are separate from the email itself). Most email is not likely to require significant file format migration in the future (given that much of the content is ASCII text), but the **attachments will require routine preservation actions to remain usable.**

- Linked content in emails references time-sensitive information (hyperlinks) which can be broken by the time email reaches the archives, reducing the fidelity of the records.

Additional Header Information

Specifications for IMF and SMTP were created to be flexible, providing only a few mandatory elements so that application developers could create interoperable software. During the expansion and development of email technology and infrastructure, as the specifications evolved, and as the number of software applications grew, developers of those individual applications made extensive use of custom email headers to enable additional functionality and use cases.

These custom and extended email headers make the typical, modern electronic message considerably larger and more complex than might be implied by the IMF specification. Since these headers are customized, they can convey any kind of information that could be relevant to the sender or recipient mail server and email client applications. Some headers might interact with mailing list subscriptions, while others might instruct readers to NOT display emails as a thread, while others might be used for message tracking and tracing options. Usage of custom headers is limited only by the choices of the email application developers.

Managing Attachments

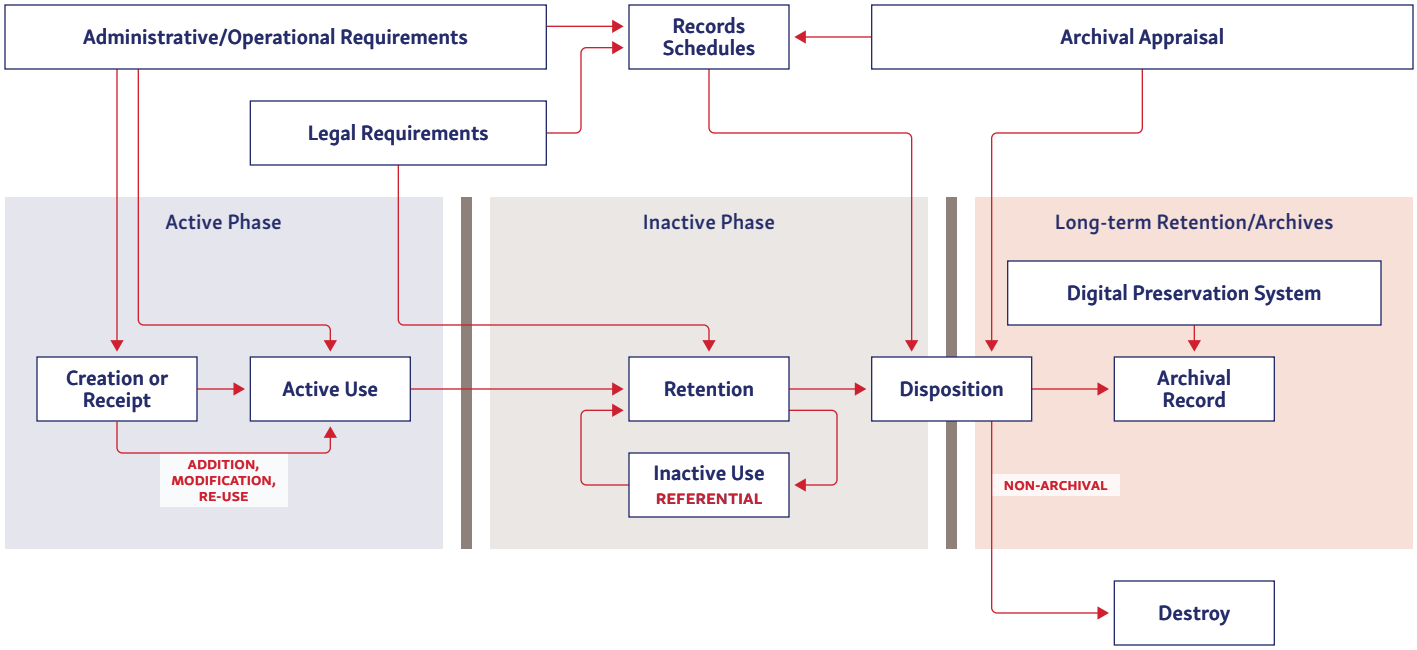
In theory, managing attachments as individual objects would not be challenging to archivists; after all, digital archivists should now expect to manage unstructured documents of various file formats and unknown provenance in the course of everyday government business. Attachments themselves rarely pose difficult preservation challenges; the process for managing attachments would no doubt be similar to the same digital preservation processes used for records throughout state and territorial governments. The difficulty in managing attachments comes in preserving the contextual linkages between the attached files and the email itself, especially during format migration and discovery processes.

The difficulty in managing attachments comes in preserving the contextual linkages between the attached files and the email itself, especially during format migration and discovery processes.

States and territories should seek OAIS-compliant digital repositories that provide tools which can handle the ingest of both emails and associated attachments as a single semantic unit. In the absence of such tools, however, there are some mechanisms which can provide added clarity and security in preserving digital provenance.

First are the email headers, which should provide information (about the MIME-type, at least) of attachments included with the email. However, this trail of digital provenance usually only goes one way, with the contextual information being embedded in the *email message* but not the attachment. (Thus, if the two are separated, the email will provide a digital trail indicating the presence of an attachment, but the attachment will provide no such evidence, linking itself back

Typical Information Lifecycle



A Note on Legacy Email Conversion

Legacy email (more than 5-10 years old) has been retained by a number of states/territories in a variety of different formats. Many email legacy exporters and transformers, especially for more uncommon software options that may have been used in the 1990s or 2000s, provided exports into common textual formats that could be read by other programs. Most common among these formats are likely raw TXT files, HTML files, Word documents, and PDFs.

Unfortunately, in many cases, the converters stripped much of the header information from the email (typically To/From/Subject/CC/BCC/Date/Time/Body are all that remain). Most modern email software (including some processing software) will not recognize these files as emails - the structure and content of the email has been lost if the email headers have not been retained. Furthermore, there are few, if any, commercial products that can convert these files back into a more suitable format (such as MBOX or EML) should that desire exist.

For this reason, email files converted to these formats without preservation of structural and administrative metadata have made the management of such files much more similar to the management of unstructured documents as are typically found in those file formats. Once data is lost, however, it is unlikely this information can be meaningfully recovered without access to the original email server/application which contains that information.

Therefore, PREPARE recommends close examination of format migration mechanisms; if there is doubt, retaining email in its native format is better than the loss of essential metadata. A bad conversion is worse than no conversion at all.

to its carrier.) Caution, therefore, must be exercised in separating attachments from their messages.

One technique that could be used is embedding some of the email header metadata (such as the unique message-id, sender, recipient and timestamp—most of which are required for transmission, anyway) inside the attachment. This way, if the attachment is ever separated from the source email, the user will be able to identify that it was sent through email, and will be able to trace it back to the original message (if preserved). However, this may be both laborious (as there are few, if any, automated mechanisms that can perform the task of writing these headers into attachments), and probably impossible in some scenarios (given that some file formats that could be transmitted through email may not have the structure or means to include contextual metadata). However, some email destination formats (such as PDF) can effectively register and make use of custom metadata such as this.

PREPARE's recommended solution is to treat the message *and* attachment as a single semantic unit worthy of preservation, and to preserve links between the two as much as possible. For instance, an archival information package consisting of multiple files (an EML and Word document, for instance) could faithfully represent the message, its attachment, and context, and could be meaningfully described using the email headers as key metadata elements in addition to any descriptive or administrative metadata about the attachment itself.

Migrating Email from Native Applications

At some point in the email lifecycle, it will become imperative to migrate email from its native application and move it to a more suitable storage format or location, either due to lifecycle (use, reference, retention, and so forth) or preservation requirements. Unfortunately, due to the highly customized way that most email applications have been configured, export and migration pathways from email applications are not always reliable, especially when utilizing third-party software.

1. **Not all email headers may be migrated:** Third-party migration tools may be optimized for use with specific software, or even specific software versions, and may not be able to keep up with the myriad use-cases across government. While public officials are likely sending email through a consistent platform, they are just as likely receiving it from numerous other platforms which may use custom headers and/or whose users may use custom fonts, embedded images, and other modifications which conversion applications might not be able to handle. **Always inspect the outputs of format converters to ensure minimal loss of data.**
2. **Attachments:** Email attachments should be embedded in the original email, and when migrating data out of the originating application, care must be taken to ensure that contextual links between an attachment and the email with which it was sent are preserved. It is recommended to use the unique message-id of the email to preserve these links whenever possible.
3. **Linked Data:** Modern email conventions include many instances of linked data, from documents hosted on external servers (instead of attachments), to embedded images and fonts which must be retrieved from an external source. Migration processes must preserve these links whenever they are meaningful. Staff conducting migrations may not have the ability to access such linked materials at the time of migration (possibly due to permissions, technical limitations, or simply that links are broken after time has passed).

Technical Recommendations

Many of the issues related to the management of email in state government originate from similar sources: phased development of email systems over a long period of time, using different software, versions, or methodologies to manage records. Email is a relatively simple technology that produces relatively simple, structured records; it is this ease-of-use that caused it to be rapidly adopted across state governments throughout the 1990's, and utilized in every context imaginable.

The technology of email has come a long way and, in some cases, has been superseded by other software: digital workspaces and collaboration platforms for document sharing/editing, chat and direct message services for communications, and others. However, email remains a reliable and familiar way for state and territorial government officials to contact, and more importantly *be contacted by* their constituents in an asynchronous way. For this reason alone, PREPARE forecasts that email will remain in use in government for some time.

Individual use-cases for legacy email collections may have issues as discussed above, and the most frequent discovery in legacy collections is the lack of email headers and other contextualizing information resulting from format conversion. Many, if not most, legacy collections will require custom solutions and workflows to address these deficiencies. Most technical recommendations, therefore, are intended to serve as day-forward approaches to managing technical aspects of email.

Capturing Externally-Hosted Content

All-in-one collaboration platforms like Microsoft Office 365 (OneDrive) and Google Workspace (Google Documents) allow users the ability to embed links to files that have a similar look and feel to attachments, but are actually links to records stored elsewhere (such as cloud storage). Some of these platforms even allow users the ability to edit documents online with their email application. Functionality like this is beneficial for the user from the perspective of productivity and collaboration, but creates preservation challenges—for example, **a link to a record hosted elsewhere that is included with an email is unlikely to survive an export out of an enterprise email system, unless specifically planned for and addressed.**

Maintenance of contextual links (including hyperlinks and other semantic linkages) will become increasingly complicated. Such emails may present as significant records that document when an individual's access to a record was established. Please see the University at Albany's Mailbag Project for an attempt to address this issue.

Capture of email from active management platforms will necessarily become more complex, and may require the fetching of attachments prior to capture, or the export of linked attachments to be handled outside of the email application. It is not recommended to transfer email and "come back for the attachments later" as links can rot and disappear, and records located at a specific location at the time of the email's sending can obviously change afterward, changing the context of the email record.

It is not recommended to transfer email and "come back for the attachments later" as links can rot and disappear, and records located at a specific location at the time of the email's sending can obviously change afterward, changing the context of the email record.

- Google Vault and Microsoft Exchange Retention Policies can restrict individual users' ability to delete emails permanently. Centralized control of email is recommended as it vastly simplifies management of email and turns something requiring extensive oversight (individual user management into email) into something local.
- Enterprise email tenants should be configured to automatically provide individual users' folder structures or labels that conform to subject matters, protected or sensitive information, casefiles, or any other organizational system that adds context and understanding. But rather than *requiring* its use (which will have low rates of compliance), providing the tools that make it easy and natural will aid in adoption.
- Automated labeling or categorization of records based on keywords, natural-language processing, or other methods are even more valuable. Most users whose emails would be retained long-term (high-level policymakers) are unlikely to devote much time to email labeling or organization without dedicated staff time.
- Extract archival email records (and their entire accounts) from active email servers once retention periods have been met. While it is in the records producer's (i.e. the public agency) custody, it should be managed in their system (and they can leverage e-discovery tools to provide access to records, if needed). But, once retention is met and records are eligible for transfer, extracting records (ultimately to a stable preservation format) for transfer to the digital archives is necessary. **Email platforms are not digital preservation systems.**
- Email can be stored in proprietary file formats for limited periods of time (<5 years), especially while the state/territory has access to the software that created it. Use of PST and possibly MSG files in the Microsoft environment is inevitable. If data storage is not a primary concern, retention of source email (.pst files) as preservation masters alongside exports is also recommended.

- Long-term preservation of email records (>5 years) is only recommended in open-source, stable formats such as MBOX, XML, or PDF/A.[14] **If styling, formatting, or external content is significant, choose destination formats that support them.**
- Preserve as much of the email header information as possible. When migrating to other formats, ensure this information is captured in the new format.
- Preserve attachments in a digital preservation repository that allows them to be individually searched as well as indexed. OAIS-compliant digital repositories have become professionally accepted standard practice in the archival community. Provide contextual links to the email and mailbox from which it originated (recommend using MessageID, plus a unique identifier for the mailbox) to uniquely identify email. Ensure that users accessing emails or attachments can easily reach the other through a trail of preserved digital provenance.

Email Management in State and Territorial Governments

Email management and preservation in government is primarily a challenge of governance. Attempts to address challenges are encumbered by legacy technology decisions and optimization for certain lifecycle elements (active use) over others (inactive and legacy management). It would be impossible to address requirements for a set of records as widely-used as email without considering preservation throughout the information lifecycle, and the management of elements within it is critical to any attempt to thoughtfully preserve information. Decisions about how to transfer and preserve email in the archives will have implications throughout its entire lifecycle.

Overview: Email Governance – Current-State

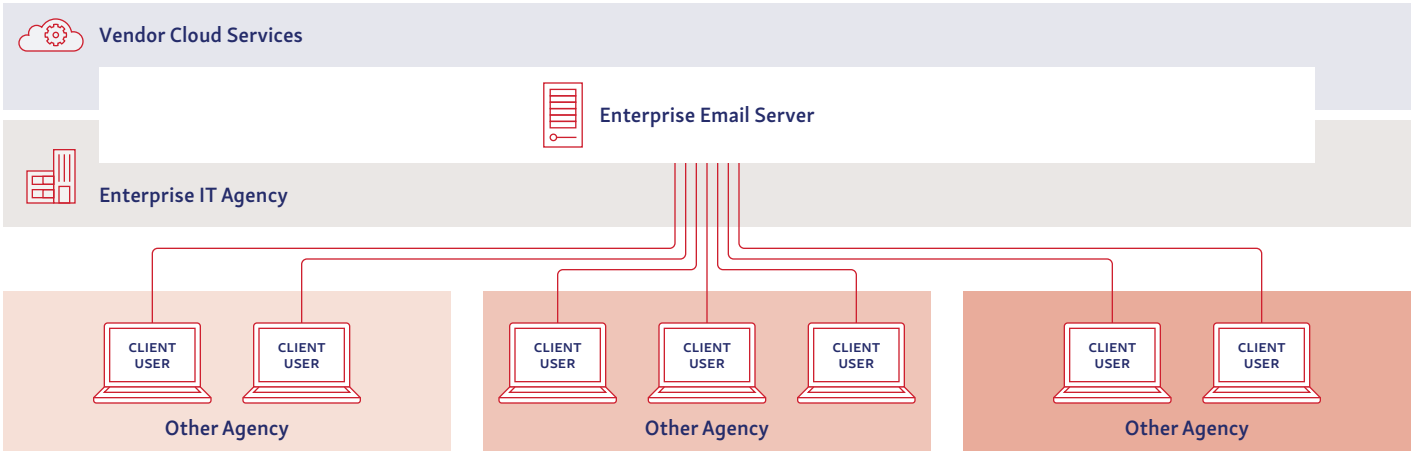
PREPARE's needs assessment survey indicated that 83% of responding states and territories utilized an enterprise email system managed by an enterprise IT agency, while a majority had at least one agency that administered its own email system (53%). Enterprise email management has become the de facto standard across state and territorial executive branches, but many of these governments have separate instances, tenants, or entirely different systems for other branches and, in some cases, specific agencies. Thus, it is highly likely that any approach to enterprise-wide email governance will need to account for these variations, or will need to be targeted specifically at executive branch agencies. Fortunately, executive branch agencies produce the vast majority of public records, and most judicial branches have modern case management systems which are equipped to ingest and manage communications as part of their model.[15]

The information lifecycle and its management are central to the effective administration of email in state and territorial governments. Email records are invaluable resources for understanding the actions and decisions made by officials at all levels, and thus offer great referential value to the government, as well as the many public communities that use public information. Furthermore, state and territorial governments are made up of thousands of employees, and those employees turn over frequently, often taking their institutional knowledge with them when they depart, and often the only evidence of their decisions is found in email records. **Email can have significant administrative value as a timestamped record of what was known at what time; tracing the decision-making process of public officials is essential to audit the processes of government.**

> *Email records are invaluable resources for understanding the actions and decisions made by officials at all levels, and thus offer great referential value to the government, as well as the many public communities that use public information.*

Unfortunately, despite its immense administrative value, many of the major obstacles to email preservation, as well as the rising costs of management, processing, and review of email can be traced back to loose classification and governance, and poor or nonexistent lifecycle management. Shared responsibilities across multiple agencies, and the ongoing need to provide on-demand access to both current and legacy email for various processes (especially FOIA and other public records access processes) add to these issues.

Common Enterprise Email Management Model



Traditional Records Lifecycle Stages

- **Creation and Use:** Records are created or received during the course of public agency business processes.
- **Retention:** Records are retained until all (legal, administrative, and other retention) requirements are satisfied.
- **Disposition:** Once retention requirements are met, records can be destroyed (if no permanent value) or transferred to archives for permanent preservation.

Preservation of email records is required throughout this lifecycle, while archives are typically most involved after disposition. Records at the archives would typically be ingested into a digital preservation system where they can be protected (from alteration or damage), described (to aid in search and discovery), migrated (to new formats and technologies, to maintain utility), and provisioned for continued access in perpetuity.

These same process steps could be applied to the governance of email, but they need to be thoughtfully assigned or gaps will be created that can threaten the flow of the lifecycle and create bottlenecks which can result in loss of data, increased costs, or inaccessibility of essential information.

The State Electronic Records Initiative, through its ongoing development of the State Electronic Records Preservation (SERP) Framework,[16] identifies the core components of an enterprise-wide digital preservation program:

1. **Identification** of essential recordkeeping requirements (to ensure that legal compliance and administrative objectives are met)
2. **Assignment** of roles and responsibilities (to ensure the appropriate staff are handling essential functions related to preservation and governance)
3. **Governance** of digital systems (to ensure they can meet essential requirements)
4. **Migrating** data (to appropriate systems, formats, or custodians)
5. **Provisioning** of access to records throughout their lifecycle (to ensure legal compliance and accountability)

While these elements are broad and simply stated, each deals with a critical element of information governance across the enterprise of government. Reviewing the current state of email governance in this area will help highlight the critical governance challenges and shed some light on pathways to improvement in email management.

Essential Recordkeeping Requirements

Recordkeeping requirements are fairly well-documented across state and territorial governments (including those for email), but existing models of classification, appraisal, and requirements gathering may not be ideally suited for the governance of email.[17] Documentation of recordkeeping requirements traditionally occurs in the records scheduling process, which each state and territory conducts using its own unique methodologies. However, most of these processes are derived from a data-driven process that utilizes inputs such as the historical organization, use, and management of records related to specific processes, or analyses of the expectations (implicit or explicit) of legal requirements including statutes, regulations, and policies.

BEST PRACTICE

Most states and territories with fundamental digital preservation capabilities have identified by statute or by policy (such as a records schedule or IT/Archives directive) that email is a format and not a record type.

States and territories have traditionally had difficulty making the transition from paper-based collections of correspondence to digital forms. In the early days of email proliferation, traditional correspondence was often conducted in state and territorial governments through official memos, while more casual correspondence was conducted over the phone. Email over time replaced both the official memo and the telephone as a primary means of communication, due to its convenience and effectiveness in asynchronous use. The migration of collections of correspondence from paper to email occurred slowly, organically developing into a hybrid model (a common theme with email usage) that states and territories see today, with email being an essential part of the fabric of government communications. Other technologies have since arisen (SMS text messages, DMs in social media platforms, chatroom-like platforms, and so on) but email remains a primary technology.

Email is both a *type* of content, similar to traditional correspondence, but it is also a *carrier* of content. Defining email as a "supertype" of formats has become more necessary to apply any kind of governance given that email is used in government recordkeeping processes in every context imaginable. Especially when considering the accounts of high-ranking officials, records related to multiple functions, and covered by multiple schedules are likely to be found, unsorted, in the same mailbox (or even in the same message).

A quick glance at just some of the common uses of email in government should demonstrate the difficulty in applying blanket governance to all email use-cases:

- Notifications from statewide service agencies (e.g. Human Resources)
- Organizational correspondence within an office
- Submissions of permit or license applications
- Delivery of automated notifications from web applications
- Regulatory oversight communications
- Discussion of high-level policymaking
- Contract submissions (bids, etc.)
- Submission of protected material as part of an administrative hearing
- Shared mailboxes for official communications

Records schedules as traditionally defined struggle to account for a content-delivery system like email, since a single message can include information related to multiple administrative processes, across multiple functions, and including information in multiple formats. These may be *suboptimal records creation processes,* but they are extremely common across state and territorial governments. Some states and territories have correspondence files stored in email servers, related to essential regulatory processes like permitting and licensing, that date back decades. Email systems are not digital preservation systems.

For these reasons, in many states, separate email management policies clarify governance of email as it pertains to records retention, and most of these correctly identify that email cannot be managed in the same way that traditional records schedules would assume; **a more holistic approach to governance is required.**

Assignment of Roles and Responsibilities

As shown in the previous graphic, roles and responsibilities related to email are extremely difficult to assign properly, due in large part to the resource-intensiveness of management. Subject-matter experts (SMEs), those who routinely use and are most educated about email, are the most frequent and common users of their own email, and given the lack of widespread governance over the means and methods of drafting, organizing, and retaining email, SMEs are almost certainly best positioned to understand the provenance and value of records within their own email accounts. But few government employees are assigned specific responsibilities related to email management, and fewer are afforded additional time to manage email to an extent that would be useful across the enterprise.

Additionally, email mailboxes tend to reflect highly-personal organizational systems as they are most often optimized for use by a single user. Use of case management and other records management systems that allow for the import of email from other systems (such as the capture of email sent to a particular mailbox) does afford some degree of redundancy to the capture and preservation of email.

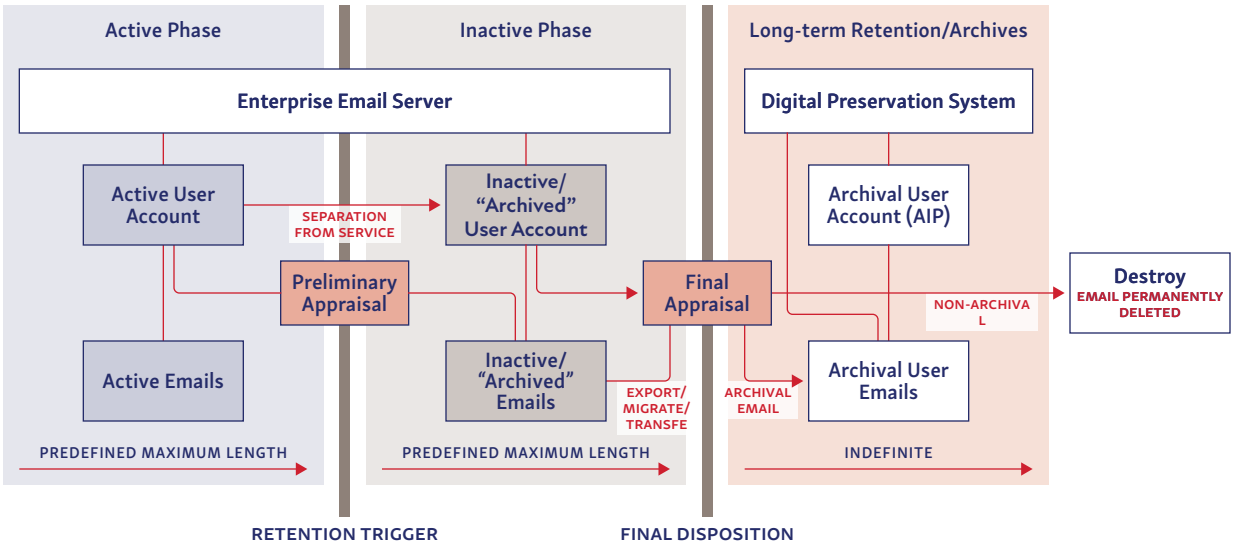
However, the timeframe in which email must be managed tends to occur after the individual has separated from government; that is, email mailboxes often first require organization after the SME has left their position (and, though some states and territories do include records management tasks as part of offboarding processes, these are few and far between, and compliance with such requirements is sporadic at best). This leads, most often, to records managers, supervisors, and oftentimes archivists as being the first or only government officials to review mailboxes to support essential processes like FOIA requests, records retention, legal discovery, and others.

Governance of Digital Systems

The mechanisms at play in modern email applications have been discussed extensively in Part 1 of *Be Prepared.* Governance of email systems falls to three major concerns: who has custody and control over records in mailboxes, who controls access to those mailboxes, and what tools are provided for those with custody to make changes (including the addition of metadata and other contextual enhancements) to records within those mailboxes.

State and territorial IT agencies typically manage enterprise (or agency) email systems, but usually the custody and control of records within is left to the individual agency user. This is fairly standard and mirrors the management of other electronic records and paper records (an illustrative example: the agency responsible for managing government buildings

Sample Email Lifecycle Manaagement Model



typically has no role in the management of records within the building, much like the IT agency typically has a limited role in the management of records within its platforms). Of course, the relationship between IT and the management of email is much more complex and varies greatly from state-to-state and territory-to-territory.

However, the configuration and default settings of the enterprise email system, as well as the way it interacts with the universe of other software applications available in state and territorial governments are significant factors at play in the management of email. For example, use of retention policies set at the server or enterprise level can restrict the ability for individual users to destroy email, or they could automatically enforce the destruction of certain kinds of email after a set time period.

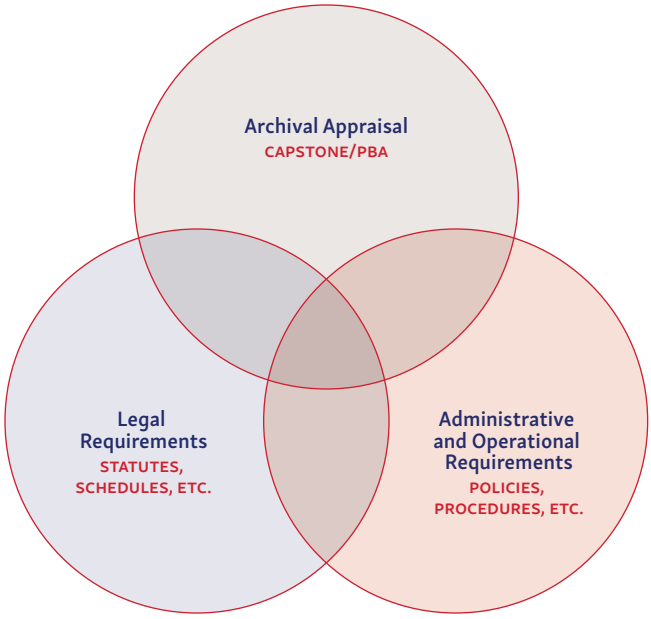
The needs assessment survey conducted in early 2021 measured not only the digital preservation capabilities of state and territorial archival agencies, but asked states to look across the government enterprise at the constellation of policies that govern the use of information technology, as specifically applied to recordkeeping. Email governance policies were included.

- **Personal Use of Email:** Many states permit the usage of email to conduct personal business, though often in limited circumstances and always worded in such a way as to protect state property and financial resources from undue burden. This is consistent with the development of email governance over the past few decades and may address a problem which is not otherwise solvable.
- **Email vs. Records Management Systems:** Email's strengths (simplicity and ubiquitousness) have caused its adoption in many business processes for which it is not suitable. Even today, despite the existence of secure upload mechanisms, records management systems with robust metadata and digital provenance controls, and case management systems that can easily handle numerous types of documents and content, email remains frequently used throughout state and territorial governments as a means of sharing information. Many of these contexts are formal (such as permit or license applications) or contain protected or sensitive information (such as court and administrative proceedings), for which the use of email is inappropriate. However, often the ease of use trumps other considerations.
- **Acceptable Use Policies:** Conversely, some jurisdictions define what information can be transmitted over email, though these are usually more specifically defined as prohibitions against use of email to transmit (usually) sensitive or statutorily protected information such as PHI and PII.

Data Migrations

Currently, there is little done to migrate email from its current state. States and territories typically manage email

Capstone captures overlapping requirements, simplifying appraisal



records within the enterprise email system. In many of these jurisdictions, email is being regularly removed from these systems due to auto-delete policies which take effect after the user separates from government service and the account is terminated; alternatively, in some jurisdictions, email resides within these systems and continues to grow in quantity.

Records management systems are not archival preservation systems, and while some of them do possess similar functionality, their fundamental goals are not perfectly aligned. Email systems are typically optimized for the delivery and receipt of email, not for complex records management tasks such as the classification of data, utilization of retention schedules, redaction, and so on. E-discovery requirements have necessitated modern email platforms to address these functionalities, but many experts on the ground consulted by PREPARE indicated that the tools available for the proper management of records in email systems do not readily exist, or are not sufficiently mature to enable the complex and oftentimes overlapping demands of email management.

Records management systems are not archival preservation systems, and while some of them do possess similar functionality, their fundamental goals are not perfectly aligned.

It is likely that email will need to be migrated from enterprise email systems (likely to new formats) if it is to be retained and preserved in any appreciable way.

Access Considerations

Access considerations have been the primary driver of what little email governance does exist. Due to the ubiquity of email and its use, email communications are often the primary target of public records requests, and the legal requirements to respond to these requests accurately and quickly have led to the modernization and proliferation of e-discovery tools that can commonly be found in state and territorial governments. These requirements, however, exist throughout the records lifecycle, and separating email from the enterprise system could lead to the inability to use these tools.

Summary of Governance Challenges

- The default email account model assumes a one-to-one, employee-to-email account relationship, but in reality many other scenarios exist, including one employee to multiple accounts, one account shared by multiple employees, and many in between.
- Similarly, email is used as a component in thousands of discrete workflows that involve permanent records across the US. Regardless of whether these processes should be using email, government archives must account for potential records contained in email for ALL government functions until otherwise indicated.
- Successful governance requires thoughtfulness about email at each level of semantic unit; different requirements may exist for individual records, sets of records, accounts, and many other possible configurations.
- Many, if not most, government mailboxes are minimally organized and described. The typical personal email account has an arbitrary organization scheme that highlights different topics or functions, organized chronologically.
- As a corollary, email is difficult to organize well. Sometimes emails cover multiple topics, contain sensitive information that needs to be separated/redacted, or needs to be referenced and used by multiple disparate processes. Depending on the user, email may not lend itself well to familiar patterns of records organization (e.g. folders).
- Access to email records cannot be interrupted without major disruptions to existing information workflows. Day-to-day work and freedom of information/public information requests necessitate access to search, retrieve, potentially review and redact email. Holds for litigation and public records requests can have short turnarounds or long timeframes during which email cannot be inaccessible.
- Modern productivity platforms integrate email functionality into their workflows, which can be a blessing if governed well and a curse if not.
- Email servers and email client applications are primarily designed to send and receive email, and are not typically optimized for traditional records management or digital preservation functions. Modern email applications can apply labels, sort email into folders, and even manage complex workflows that route, organize, and potentially delete (or mark for deletion) records, but these must be planned for and configured appropriately.

Governance Recommendations

Much of the previous section has been focused on outlining how email is difficult to manage due to its quantity, wide variety in use cases, and lack of clear requirements across state and territorial governments. Therefore, PREPARE suggests some broad steps that can be taken by any state or territory that wishes to improve email governance jurisdiction-wide. CoSA recognizes that not every suggestion will be possible for every state or territory, nor will the recommendations offered be perfect. However, we hope these recommendations can address the management of 95% or more of the email in state and territorial governments through the implementation of common-sense governance steps that should not be resource-intensive to implement; the remaining tapestry of corner-cases, non-optimal uses of email, and other difficult-to-articulate situations should be addressed in a case-by-case manner that optimizes for adherence to requirements, including retention and public inspection.

Classification and Appraisal of Email

The needs assessment survey gathered that 87% of responding states/territories have records schedules (or similar instruments and other policies) that address email specifically. In most cases, states do not discriminate between paper and electronic records, and emails are broadly covered in schedules which tend to be agnostic to the format of the record. However, the needs assessment survey identified that despite statute and policy (including records schedules) covering email, there were still bottlenecks in the email lifecycle. Why?

Email retention does not lend itself well to traditional record schedule and retention requirement models. There is unlikely to be a definition for email in a schedule that would even be applicable to all of the possible use cases described in this document. Furthermore, due to its use as a transmission medium, it's possible for a single email to have relevance to multiple processes that could be governed by multiple schedules. Traditional models of records scheduling governance are typically based around content-based appraisal; that is, the retention is determined by the content of the record, not necessarily its format or characteristics. Appraisal for email at this scale (millions of records produced and received by states and territories each year) will be resource-intensive, quite simply because there is a huge amount of data that needs to be read, analyzed, and processed, whether by human archivists or by machine processes.

A single account can easily have hundreds of thousands of messages, or more, especially for public officials. Word-by-word content-based appraisal is simply not practical *at this time* for a variety of reasons, chief among them the difficulty in parsing text across multiple file formats (attachments). Certainly, recent developments in machine-learning could mean in a short time that parsing of text could be a viable option. For this reason, other approaches have been attempted, and the most useful of them in the context of government has been a position-based appraisal approach devised by the National Archives and Records Administration called Capstone.

The premise of position-based appraisal: most email accounts are attached to individual employees during the course of their public service. Rather than attempt to deeply analyze the contents of those mailboxes, some of which may contain hundreds of thousands of messages (or more), the state/territory instead can analyze the nature of the work performed by the account owner and determine how much they are involved with functions that produce permanent records. If they are a high-ranking official likely to be involved in decisions affecting policy, then information contained in their email accounts is likely to reflect that role. Instead of analyzing millions of records, all that must be analyzed is the role and position of the user. A high-ranking official, or one that administers a major program, is very likely to have records in their email account related to the administration of that program. If records from that program are classified as permanent, then so, too, should that account be classified as permanent.[18]

While those who have employed the Capstone approach have done so in their individual ways, PREPARE has generalized this method and recommends designing this approach around practical workflow-based tiers (see chart below).

Those at the highest levels of government, including elected officials, agency heads, and their principal deputies either:

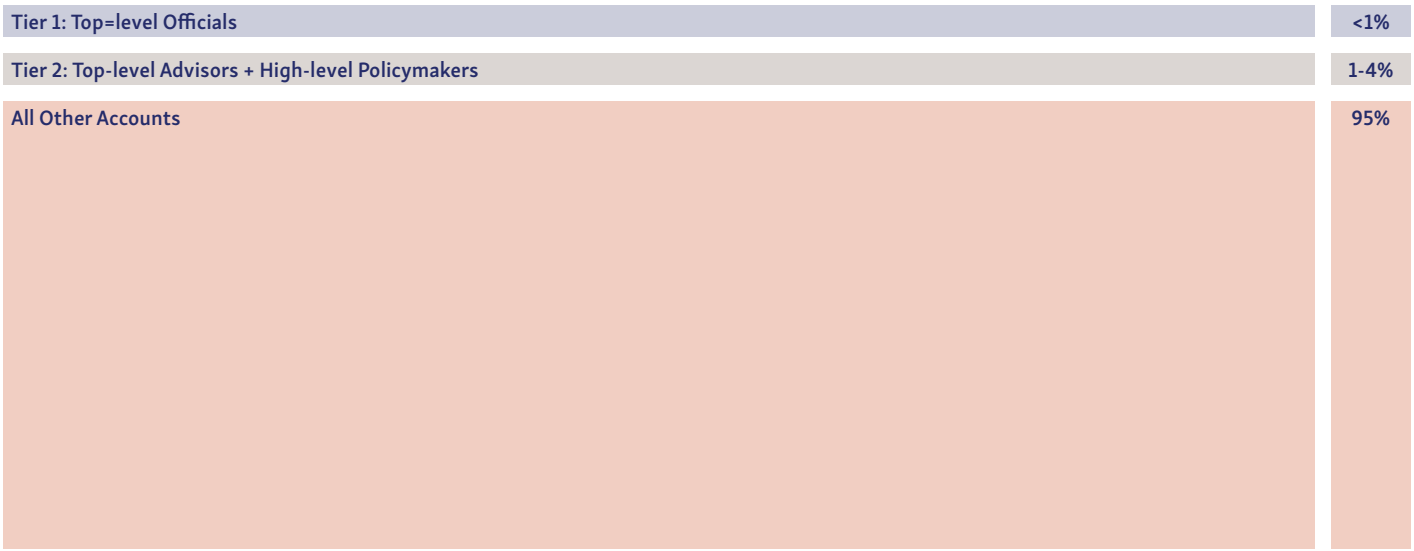
- Already have their correspondence (including email) classified by statute or schedule as permanent, or
- Have an extremely high likelihood of having records that are classified by statute or schedule as permanent.

Rather than focus resources on appraising email before it reaches the archives, these accounts can be classified as permanent at the outset of the records lifecycle. Similar, too, are low-ranking accounts that comprise the majority of the accounts issued to the state and territorial workforce: these accounts are unlikely to contain major policy-affecting work and are unlikely to need significant review (notable exceptions apply, such as transactional staff working with data classified as permanent). A third category that falls between these poles, high- and mid-ranking staff, is a gray area and will likely depend on case-by-case classification. These emails require some review and appraisal, either at the position or in the content of the mailbox.

Certainly, the state or territory that applies position-based appraisal can and should develop as many categories as needed to meet legal, administrative, operational, and preservation requirements. However, PREPARE recommends beginning with as few categories as possible and designing these classification categories around expected workflow steps. This will greatly aid in understanding this method, its adoption and execution.

One of the most challenging aspects of preserving email is determining what email to preserve, and how. Best practices found in records scheduling throughout the U.S. indicate that a content-based approach to determining retention is preferred for most (unstructured) electronic records, and for email

Position-based Email Appraisal Categories + Estimated Percentage of Accounts



the same is true. As discussed, since email can contain any type of content, preservationists must assume that, without strictly-controlled governance, it does contain those kinds of records. Determining the long-term or permanent archival value of email, therefore, is necessarily complicated by needing to review many details of an email: headers, body, attachments, and context, not all of which may be readily available.

BEST PRACTICE

Whether using position-based appraisal or not, some user mailboxes will contain email that must be retained for longer time frames than other email in that same mailbox. For example, correspondence that comprises part of a casefile may need to be retained as long as the case is open. PREPARE recommends developing workflows to migrate or copy these records out of the email system and into a records management system that is more suited for this work. For example, all emails related to a casefile can and should be integrated with the rest of the case records in a case management system.

Content analysis performed on a set of records with the scale of email (potentially hundreds of thousands of records, or more, in a single mailbox) is extremely difficult without the aid of information-processing tools. Even with advanced processing tools,[19] it will be difficult to apply them across the entire enterprise, given the significant variations even within one program, office, or individual official's use of email.

Furthermore, most of the analysis happens without ever needing to read and comprehend the more granular semantic units. Whether a program director has essential information in their email largely depends on *how the system was used* and *what other systems for storing essential information* are in use and capturing that official's work. In some agencies, for example, directors may never use their email accounts for policy development or essential decision-making, as those records may be captured in casefiles or formalized proceedings that are already captured outside of the email system. In other cases, such as those agencies with lower information governance maturity, email may be the *only* record of certain kinds of high-level administrative work that would be deemed permanent in most jurisdictions.

Ultimately, position-based or Capstone appraisal methods collapse most of the appraisal work to two key parts of the email lifecycle workflow—a periodic review of positions to determine the likelihood of permanent records appearing in an account, and a periodic review of "Appraise" accounts at the end of their inactive life. Consolidating this work to key times reduces the "always growing, never shrinking" dilemma of email, and furthermore can make the workflow much more manageable by allowing agencies to shift resources at appropriate times.

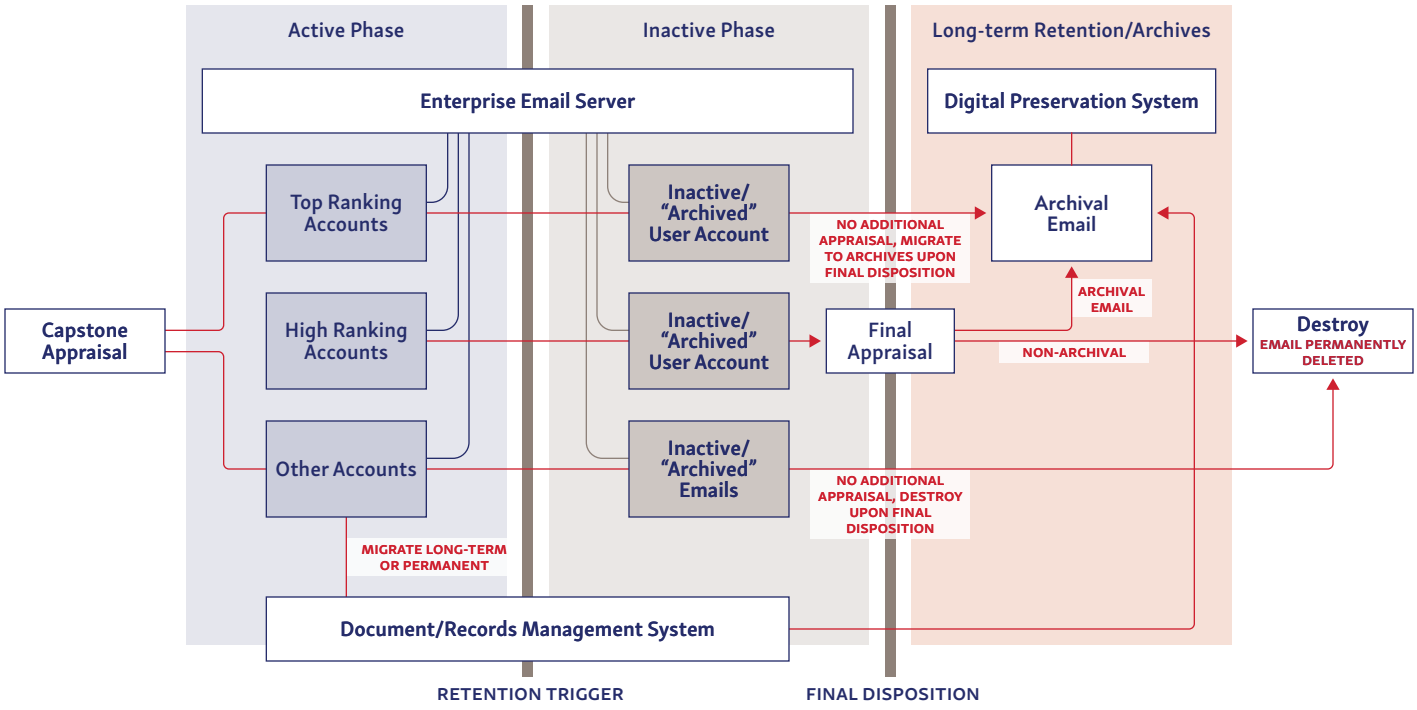
Note that Capstone or position-based appraisal methods attempt to address the problems of retention and archival appraisal *only*. These approaches do not address the identification of protected, exempt, or otherwise sensitive material which could obviously appear in any account regardless of how it is appraised. Furthermore, this appraisal approach does assist in identifying those critical records which must be retained, and for which the standards of description and preservation must be higher; it does not assist with mechanisms like FOIA and similar laws, which require the analysis of content at a much more granular level to support public access requests. The position-based approach is best utilized to quickly determine which accounts must be retained indefinitely.

PREPARE strongly recommends the utilization of this approach, especially for those states and territories that are just beginning to implement policy frameworks governing the management, retention, and preservation of email. Utilizing this approach will greatly simplify the work that must be performed by archivists and records managers, allowing them to evaluate and adjudicate thousands of positions easily. Most importantly, for the most high-value email sets (such as emails of governors and other elected officials) the position-based appraisal approach does not preclude other methods of appraisal or processing.

Other mechanisms will need to be used to determine, for instance, how long non-Capstone accounts should be preserved and retained; this will depend greatly on the individual statutes and regulations in place in each jurisdiction. To accomplish this, PREPARE recommends considering the lifecycle of email in three specific phases (see table below).

LIFECYCLE STAGE	DEFINITION	LENGTH OF PHASE	POSSIBLE LIFECYCLE PARAMETERS
Active	Email that is actively used and frequently referenced as part of ongoing business processes.	Short (3-5 years)	All active staff members using email accounts.
Inactive	Email that is no longer actively used and is infrequently referenced, but occasionally useful.	Medium (3-5 additional years)	Email is moved to online/offline archive but still accessible to user. All separated staff email accounts immediately become inactive.
Long Term Retention and Disposition	Transition to final lifecycle stage; permanent records are migrated out of enterprise email system and non-permanent records that have met retention are destroyed.	Immediate (destruction) or Indefinite (permanent preservation)	Inactive email that has met its retention and other legal requirements can be destroyed if not archival; or migrated out of enterprise systems if the agency wishes to retain it for reference purposes.

Sample Capstone Email Lifecycle Management Model



Use a position-based appraisal system for classifying email, rather than a record-by-record approach. Position-based appraisal systems like Capstone have proven successful in both federal and state governments.

- Workflow steps that involve the review and/or redaction of protected, sensitive, or exempt information should be conducted by the records producer or current custodian (not IT or the archives). These processes are resource-intensive and often require detailed legal or subject-matter expertise, and therefore are most effectively accomplished where that expertise resides.
- Similarly, users are best positioned to categorize, label, and organize their email records. If this is not done, it is likely more appropriate and efficient to preserve original order rather than have another staff member (such as a records manager) weed or reorganize records prior to transfer.
- Leverage position-based appraisal to reduce the amount of resources spent on managing mailboxes not requiring long-term retention. Do not plan to process individual messages except in high-profile or legally-mandated cases.
- Enterprise IT agencies should prohibit, through policy or technical measures (such as retention policies in email clients) individual users from permanently deleting email records. There is little value in the user performing these tasks (which are rarely done, anyhow) and run the significant risk of loss, decontextualization, or improper behavior. Individual users are unlikely to be aware of the broader records management context for records.
- Manage accounts or mailboxes instead of messages. At the enterprise level, managing individual records is resource-intensive with little gain.
- Utilize automated labeling, organization, and classification methods to reduce overhead for email management. Pre-program options for the above in enterprise email systems to encourage users to use them.
- It is probably not possible to prevent, by technology or policy, personal use of email by individual employees. At best, policies against this use do curtail incidence rates, but it is unlikely to ever eliminate it. This is especially confounding when concerning the records of elected and other political officials, who occasionally have a poorly-defined boundary between their public service and personal or campaign information spheres. PREPARE recommends that, no matter the policy on personal use, officials should clearly understand that:
  - All emails in PBA-appraised accounts will be retained. The official cannot delete these records.
  - Personal records, if not identified by the user, are unlikely to be found by records officers or archivists.
- For records sent or received by email that have retention requirements that extend beyond 7 years, PREPARE recommends exporting and retaining those records in a system more suitable for long-term preservation. Migration of records out of the enterprise email system and into a content-management or digital preservation system is strongly recommended.

Records Management and Archival Workflows

Overview: Preserving Email

States and territories indicated during the needs assessment and the focused testing phase that greater understanding of basic processes is needed in order to preserve email. The issues of scale confound attempts to begin preserving email, as simply getting started can be the biggest challenge.

Persistent, continuous, useful access to email will be defined by the preservation actions that states and territories take to capture the content and context of email throughout its lifecycle. In many ways, preservation of email is similar to the preservation of other records and follows a familiar preservation pathway:

- Capture content and context (at the time of transfer)
- (Generate and) verify fixity
- Weed out malicious (virus/malware) and non-record material
- Determine format migrations (if any) or normalizations
- Accession into digital preservation repository (with all associated links)
- Create/enhance descriptive information (including metadata)
- Plan for long-term preservation monitoring

PREPARE offers some sample workflows (and process steps) that could be modified and adopted to a wide variety of scenarios. If implementing new workflows, some basic guidelines apply:

1. **Start small, then expand:** Test processes for success, pilot processes, and then refine before expanding. There is no need to start with a massive project.
2. **Iterate:** Every process will be rough when it is first envisioned, and will be polished over time. Don't expect to get everything right the first time.
3. **Focus on preservation:** Preservation actions and decisions are difficult to undo and can have lasting consequences.

PREPARE focuses on inputs and outputs for its sample workflows so that they can be inserted (with some customization and modifications) into pre-existing frameworks as long as the inputs exist. Outputs of each workflow are essential, as they allow archivists the ability to move to the next workflow (e.g. preservation actions) as necessary.

As part of this section, we share some basic workflows for utilizing some PREPARE-tested software in email processing. These resources are not nearly as comprehensive of the work done by the developers of each of these tools, so it is recommended to check the project site for documentation that more thoroughly instructs how, and when, to utilize each of these.

Email Preservation Challenges

CHALLENGE 1:
Selecting Significant Information to Migrate

No matter the appraisal or management model, at some point in the email lifecycle, some data will need to be exported or migrated from the enterprise email server(s). Optimizing the timing for when this occurs can offer some resource savings, but the process of export (and what to do afterward) is far from simple.

Firstly, there are few options when it comes to the most ubiquitous email platform in state and territorial government, Microsoft Exchange/Outlook, either on-premises or in the cloud. Microsoft's email systems are very advanced and offer tremendous functionality that goes far beyond the simple requirements outlined in the IMF specifications. Calendars, task lists, contacts, and other contextual data make up a significant portion of the information associated with a user's account, and when exporting data from these systems, most of this information could be brought with it. As far as PREPARE knows, there is no native functionality in Office365 or Exchange to export email messages from an account to MBOX, EML, XML or any other open-source format. PST and OST, Microsoft's proprietary data formats, are the destination formats allowed; third-party add-ons and tools abound that offer the ability to convert PST to MBOX and other formats.

Recommendation: The ideal format migration results in no signal loss; for email, this means a faithful preservation of all headers, body, and attachment(s). Thus, when using a format converter, inspect migrated files to ensure that, at minimum, the following critical header information is retained:

- **Sender information:** local-part (account) and domain-part (server). Example: local@domain.com
- **Relationship information:** message-id and references
- **Digital provenance:** trace fields (includes server time/date stamps)
- **Subject**
- **Body**
- **Attachment(s)**

CHALLENGE 2: To where and what are we migrating

For this reason, use of open-source, flexible formats like EML, MBOX, or XML, which can accept all of the relevant information, are recommended. However, as has been discussed, many converters and exporters of this type are unreliable when it comes to preserving all the headers, especially those custom headers that are not typically expected from certain email applications.

Recommendation: Using EML and MBOX.

- Migrating to EML and MBOX provide the most long-term flexibility. They are open-source formats and are very loosely defined, and are usable by most modern email applications, including email format converters.
- Care will need to be given to storing attachments (which will not be embedded in these files). Contextual links from the message to the attachment must be created by the conversion/export program and those must be preserved.
- Using EML or MBOX files in conjunction with a preservation archive format like that specified in the BagIT specification or an extension of it like, (another EA:BCC project), Mailbag.

Recommendation: Using EA-PDF.

- At the time of writing, EA-PDF is still in development, but offers the best of both worlds: preservation of the original email object, in its native format, embedded within an archival PDF/A, with the metadata from that record stored in the EA-PDF as well. The archival community is likely years away from the finalization of the specification and the development of tools compliant with the specification, so immediate utilization will require the use of PDF/A.
- The only drawback of using this format might be its size, and possible limitations on the number of email applications that will allow ingest of these records in the future.

The needs assessment measured significant interest in developing strong email management policies, as well as managing the preservation activities required by email records. CoSA noted that appraisal and processing are areas in which state and territorial archives need support, as 17% percent of respondents indicated that processing is an area of concern.

Archival processing in the context of email refers specifically to: the set of activities that turn a "raw" set of inactive records into an archival information package (as described by the OAIS model). Specifically, management of active records (including labeling and organizing) is not included within the scope of traditional archival processing. Unfortunately, many actions and activities that *should have* occurred during the active/inactive lifecycle phase must be retroactively applied in order to fully utilize email as an archival records set.

Digital preservation actions (such as file format migration) can (and in some cases should) occur throughout the records lifecycle, or they could occur far into the future after ingest into a digital repository.

Email Preservation and Processing Recommendations

Levels of Preservation

Not all email accounts need to be handled the same way by the archives. Some accounts may contain highly valuable information of public interest (such as a governor's correspondence), while others may contain information that is valuable in the aggregate rather than the individual. If, as described in the position-based appraisal approach, something like 2-5% of email accounts are retained permanently, it is unlikely that all of them can be processed to the same degree.

PREPARE recommends two levels of "processing attention" depending on the nature, value, and expected use of the records, which can best be determined by the archives:

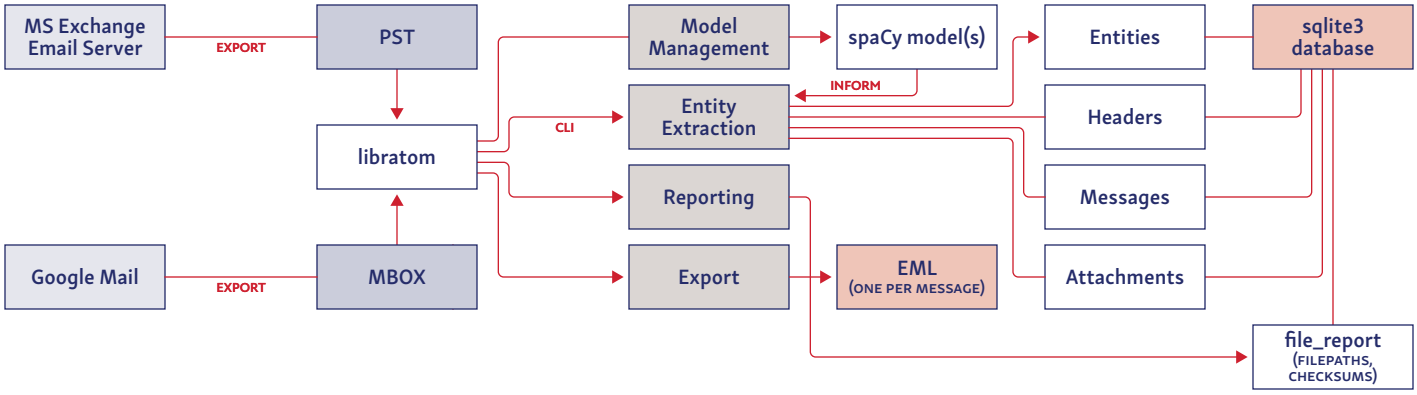
Routine (standard of practice):

- Preservation of original bitstreams, metadata, and content (email and attachments)
- Scanning, labeling, and separation of restricted material
- Mailbox or folder-level arrangement and description of records
- Preserved in digital repository as a single semantic unit (e.g. MBOX) or archive-format package (users can export and manipulate with their own tools)

Premium (reserved for high-value, visible, or significant accounts):

- Preservation of original bitstreams, metadata, and context
- Migration of messages to stable preservation formats
- Migration of attachments to stable preservation formats
- Thorough item-level arrangement, description, and processing (probably utilizing some degree of natural-language processing or other artificial-intelligence enhanced tools)
- Thorough scanning for protected information, and creation of a separate "restricted file" containing all such messages (until exemption timeframes expire)
- Extraction of entities including senders, recipients, subject matters, format types, and more
- Item-level search and retrieval, enabled by controlled vocabularies of entities (as described above)
- Custom access tools or platforms that leverage and enable specific search, retrieval, and collation of individual messages or custom queries (one such example could be to use ePADD's discovery module to provide access)

Sample libratom workflow to extract information



Tools for Recommended Use

There are many tools capable of providing some of the services required to manage email throughout its lifecycle. Overviews of some of those tools, and their possible use cases, are provided below. During the testing phase, digital archivists and technologists struggled more with knowing when and why to use specific tools, rather than how to use them.

Many of the tools listed here have comprehensive user guides which address the finer points of using that software. Instead of rehashing or summarizing that expertly-crafted information, PREPARE focuses on *inputs and outputs*.

Deciding whether to use these tools in existing workflows, or whether to build new workflows around them, depends heavily on the kind of products that they produce and what inputs are required. For example, DArcMail requires specifically-formatted MBOX files to function properly; big variations in the structure or syntax of the MBOX will cause the software to throw errors, and use of another format (such as PST) would require another tool to convert.

For a full discussion of the tools used and testing process, please see: PREPARE Phase 2: Tools Testing.

libratom Libratom is a command-line tool developed by the University of North Carolina and the State Archives of North Carolina to use natural-language processing tools to parse emails when looking for information. It can also convert a PST into multiple constituent emails for long-term preservation (EML format).

At-a-glance: Command-line software that leverages natural-language processing to process and report on email

- **Inputs:**
  - PST or MBOX
- **Outputs:**
  - Migration of data to EML
  - Extraction of entities to sqlite3 database
  - Capture of technical metadata to sqlite3 database
- **Recommended uses:**
  - Capture of metadata (entities and technical)
  - Preservation migration (to EML)

DArcMail DArcMail is an email processing application designed and used by the Smithsonian Institution to manage large email collections.

At a glance: Open-source software licensed under the MIT license (and is built off open-source technologies)

- Optimized for accounts with up to 250,000 messages
- **Inputs:**
  - MBOX (required)
- **Outputs:**
  - XML (compliant with Email Account XML Schema, EAXS)[20]
- **Recommended uses:**
  - Creation of finding aids
  - Identification of protected information
  - Preservation migration
  - Access

ePadd ePADD (now ePADD+) is an email processing application now stewarded by Stanford University, Harvard University, and the University of Manchester.

At a glance: Open-source software licensed under the MIT license (and is built off open-source technologies)

- Optimized for accounts with up to 250,000 messages
- **Inputs:**
  - MBOX (required)
- **Outputs:**
  - XML (compliant with Email Account XML Schema, EAXS)[21]
- **Recommended uses:**
  - Creation of finding aids
  - Identification of protected information
  - Preservation migration
  - Access

Key Email Management Workflow Steps

CATEGORY	STEP	SAMPLE ASSIGNMENT	NOTES / RECOMMENDATIONS
Email Management	Preserve email attachments	All	Emails are frequently used to transmit other records. Attachments are essential context to understanding the message, and therefore must be preserved according to best practices.
	Preserve email bitstreams (body, headers, other metadata)	All	All records custodians are responsible for the preservation of email and all associated contextual or format information.
	Create a derivative "copy" for agency reference	Agency + IT	Any derivative should be clearly labeled as such, especially for permanent records, to prevent duplication (and additional future appraisal)
	Administration of email server infrastructure	IT (including agency IT)	In most states and territories, IT agencies (or agency-specific IT staff) are responsible for administering email servers, including determining configurations, access, and potentially more. Any decisions related to the management of email should be made in coordination with IT staff.
	Develop jurisdiction-wide email governance framework	All	Collaboratively (agencies, IT, and archives are all major stakeholders) determine email governance framework, including roles and responsibilities related to email management.
Appraisal and Selection	Define active / inactive periods for "general" email	Agency + Archives + IT	Many states and territories may wish to establish consistent minimum and maximum timeframes for the management of email in enterprise systems. PREPARE recommends clear delineation between retention of email and management within the enterprise email system. Retention requirements are typically agnostic of system.
	Establish "Capstone" classification categories	Archives	If utilizing Capstone or position-based appraisal, determine selection criteria for Capstone officials (such as: mandated by statute, classification by pay grade, job title, or position code, and so on)
	Classify positions according to "Capstone' classification	Agency + Archives	Agencies and archives should collaborate to classify positions according to predetermined categories
Migration	Export Accounts	IT	Extracting all email-adjacent records (emails, folders, calendars, and other email-adjacent items)
	Export Email(s)	IT	Extracting individual messages from enterprise servers; export to offline (non-server) email formats such as MBOX and PST
	Transfer contextual information	Agency + IT	Email accounts may require additional contextualization (e.g. who regularly used a shared account, and when) to retain provenance; this should be collected and transfered at the time of migration.
	Transfer emails from one custodian to another	Agency + IT	Create documentation to establish chain-of-custody during transfer process, which becomes an essential part of the accession record.
Archival Processing	Survey and data collection	Archives	Archivists (or records producers) may wish to survey accounts for senders, recipients, content (such as protected information), attachments, and others. Raw statistics about message count and size can impact preservation and transfer decisions.
	Weed non-permanent emails	Archives	Archivists may choose to delete messages that do not meet appraisal criteria; all removed data should be carefully logged
	Fixity creation / verification	Archives + IT	PREPARE recommends creation of digital object fixity prior to custody/systems migrations due to the risk of loss during transit; verify fixity as part of migration to each new storage environment or repository.
	Arrangement of messages, attachments, and other email data into semantic groups	Archives	Archivists may prefer to perform post-processing arrangement in certain cases; such may be required to protect sensitive information (such as migration of restricted messages)
	Description	Archives	Archivists may choose to provide additional descriptive information about the account holder, context, and so on. If an account is presented as a single file (e.g. MBOX), descriptive metadata could include common correspondents, subject matters, and more.
Access	FOIA and other public records requests	All	Custodians must provide access to records as required by law. Migration, processing, and other workflow steps should not needlessly inhibit access processes for extended timeframes. These processes can occur at any stage of the information lifecycle and are necessarily mediated.
	Open access	Archives	Email collections (accounts, messages, etc.) that have been fully processed, and have protected information redacted or separated, should not require mediated access.

CONCLUSIONS

It's been close to two decades since technologists, preservationists, and pundits began to declare that email, as a format and as a medium, was near the end of its life. As governments move forward, email continues to be used as much as, if not more than, any other time in history.

There are other platforms being used for communications that have threatened email's throne: press releases or other public statements are often delivered through social media platforms like Twitter, Facebook, and Instagram. Messenger platforms like Microsoft Teams and Google Hangouts/Meet have become ubiquitous in state and territorial governments, shifting some of the more "casual" conversations away from email to those platforms. File-sharing services like Google Drive, Microsoft SharePoint and OneDrive, and Dropbox have become extremely common. And, as more of these platforms become integrated into all-in-one content management systems, perhaps some of the issues with the "generally-issued" email account service model will begin to recede.

One thing that remains true in government is the necessity for officials to communicate internally, as well as externally with members of the public; email has provided, and continues to offer, a simple, lightweight, and familiar platform for doing just that. And as long as that is true, its use will continue in government until similar functionality is replaced.

APPENDIX 1

Advocacy Talking Points: Building Capacity Through Building Relationships

Management and governance of email is a necessarily collaborative endeavor that requires the coordination and cooperation of numerous state/territorial agencies. At a minimum, preservation of email requires the collaboration between the enterprise IT agency, records producing agencies, and the state/territorial archives, not to mention other agencies (such as HR, etc.) that are consumers of email and providers of contextual information.

To aid in moving email governance conversations beyond "there's too much" and "whose problem is this" or "can't we just delete it all?", find below a few talking points that will help begin the process of relationship building and collaboration.

Governance

- *Email of high-ranking officials is archival in nature.* Most states and territories have records schedules which define administrative records as having permanent value.
- *Email is difficult to govern because it means different things in different contexts.* It's a tool, a platform, or means of transmitting information, rather than a specific type of record. Furthermore, almost all state and territorial employees use email daily, producing a large quantity of records, and all use them in different contexts simultaneously.
- *Challenges of email management are in governance more than technology.* Governments need to move toward shared governance among agency users and records managers, information technology system administrators, and archivists and preservationists.
- *Transparency, documentation, and clear assignment of responsibility* are all extremely valuable tools for communicating requirements to users at different levels of authority and technical expertise.
- *Control of email deletion is best centralized at the enterprise email server.* There is great risk, and little overall value, to allowing end-users to permanently destroy email.
- *Loose governance of how email is used can make management challenging:*
  - Essential and inessential records are mixed in the same "box."
  - Some processes use email when other platforms are more optimized for the task (such as a form submission workflow in a content-management system)
- *Email management challenges are most acute at bottlenecks.* Bottlenecks occur where analysis needs to happen.
  - **FOIA/Public Information Requests:** Requires attorney or subject-matter expert review of records
  - **Archival Appraisal:** Determining which emails are to be retained
  - **Audits/Holds/Other Legal Processes:** Holds on destruction, alteration, organization, or other typical email lifecycle management processes
- *Guidance distributed to users must be simple, succinct, and easily referenced.* No matter how much guidance is given, it is likely that *automated mechanisms* will be far more effective.
- *Those officials with the most valuable email*—elected officials, agency and department heads, and other high-ranking policy-affecting officials—will have little time to spend organizing or classifying their email.
- *Stricter information control policies (what information can be transmitted by email) can reduce the overall quantity of email,* quantity of sensitive or protected information transmitted by email, and search and retrieval time.

Costs

- *An ounce of prevention is worth a pound of cure:* costs to undo records management mistakes are much greater than implementing workflows that reduce or eliminate errors.
- *Decisions made early in the information lifecycle have significant effects downstream.* How end-users organize, label, and classify email makes a big impact on how resource-intensive it is to appraise, search, and process.
- *Email systems have lots of inertia in government. They are big systems that affect everyone.* There will be some cost to making significant changes to processes that may not be realized immediately.
- *Scanning for protected information is a laborious, resource-intensive process.* Investments in appraisal and processing methods, or classification and labeling, to enable rapid analysis and processing of email records in a variety of workflows (public records requests chief among them) will save resources.
- *Reducing the time spent on public records requests will reduce those costs significantly.* The federal government spent roughly $1 billion on FOIA costs in FY20-21, and these were the highest overall costs since the data has been collected.[22]

APPENDIX 2

Bibliography: References, Resources, and Further Reading

Below find some useful resources for the management of email throughout its lifecycle.

Standards and Specifications

- RFC 733: Standard for the Format of ARPA Network Text Messages, November 1977.
- RFC 822: Standard for the Format of ARPA Network Text Messages. August 1982.
- RFC 2822: Internet Message Format. April 2001.
- RFC 5321, Simple Mail Transfer Protocol (SMTP), October 2008.
- RFC 5322, Internet Message Format. October 2008.
- RFC 6376: DomainKeys Identified Mail. January 2020.
- RFC 6532: Internationalized Email Headers. February 2012.
- RFC 6854, Update to IMF to Allow Group Syntax. March 2013.
- RFC 8493, BagIT File Packaging Format. October 2018.
- OMA Multimedia Messaging Service (MMS). September 2011.

Multipurpose Internet Mail Extensions (MIME)

- RFC 2045: MIME Part One: Format of Internet Message Bodies
- RFC 2046: MIME Part Two: Media Types
- RFC 2047: MIME Part Three: Message Header Extensions for Non-ASCII Text
- RFC 4289: MIME Part Four: Registration Procedures
- RFC 2049: MIME Part Five: Conformance Criteria and Examples
- RFC 6838: Media Type Specifications and Registration Procedures
- IANA MIME Type Registry: https://www.iana.org/assignments/media-types/media-types.xhtml
- Library of Congress, Sustainability of Digital Formats, Internet Message Format.

Research, Reports, and Tools

- Email Archives: Building Capacity and Community. University of Illinois and the Andrew W. Mellon Foundation. https://emailarchivesgrant.library.illinois.edu/
- InSPECT (Investigating Significant Properties of Electronic Content) Final Report. JISC, 2007. Retrieved from: https://significantproperties.kdl.kcl.ac.uk/methodology.html
- Persistent Digital Archives and Library System research project (PeDALS) Email Extractor. (Latest update: 2015)
- Prom, Christopher J. Preserving Email, 2nd Edition: DPC Technology Watch Report. May 2019. Digital Preservation Coalition. http://doi.org/10.7207/twr19-01
- Task Force on Technical Approaches to Email Archiving. The Future of Email Archives, August 2018. https://www.clir.org/pubs/reports/pub175/
- Smithsonian Institution, DArcMail https://siarchives.si.edu/what-we-do/digital-curation/email-preservation-darcmail
- State Archives of North Carolina, Transforming Online Mail with Embedded Semantics (TOMES)
- University of North Carolina and State Archives of North Carolina. Review, Appraisal, and Triage of Mail (RATOM).
- Stanford University Libraries, ePADD. (also included in EA-BCC Phase 4)
  - https://library.stanford.edu/projects/epadd
  - https://github.com/ePADD/epadd/releases/tag/v9.0 (Latest update: Version 9, October 2022)

Email Processing Projects

- Kaine Email Project: https://www.virginiamemory.com/collections/kaine/
- Using Predictive Coding to Process Email Collections: https://www.statearchivists.org/blogs/cosanews1/2019/12/04/processing-and-preserving-governors-office-emails
- State of Washington Email Management Guidance: https://www2.sos.wa.gov/archives/recordsmanagement/managing-emails.aspx

Other PREPARE Resources

- PREPARE Phase 1: https://www.statearchivists.org/viewdocument/cosa-prepare-needs-assessment-surv
- PREPARE Phase 2: https://www.statearchivists.org/viewdocument/cosa-prepare-phase-2-testing-repor
- PREPARE: Choosing an Email Appraisal Approach
- PREPARE: Developing Email Preservation Policies
- PREPARE: Email Preservation Packages
- PREPARE: Email Roles and Responsibilities

ENDNOTES

1 For an overview of the EA:BCC program, see: Email Archives: Building Capacity and Community Video Overview

2 https://www.clir.org/pubs/reports/pub175/

3 RFC 5321: Simple Mail Transfer Protocol

4 RFC 5322: Internet Message Format. See for discussion of the context of electronic mail and envelope/contents metaphor.

5 RFC 2045: Multipurpose Internet Mail Extensions expanded the definitions of what could be included in message bodies, including non-ASCII encoding in bodies and headers, multi-part message bodies, and most importantly, a specification for different formats to serve as message bodies (i.e. attachments)

6 RFC 5322: Internet Message Format.

7 A central registry of IMF header fields was created by RFC 3864 and is tracked at RFC 4021. Many, if not most, of the header fields registered here reflect legacy practices and are not intended for general use. It is possible that the registry can provide insight into legacy email collections and the software used to develop them, and is provided for reference purposes.

8 Note that in many legacy email conversion (and export) applications, some of the message headers may have been stripped. Most often this is the consequence of inconsistent or misapplied header field mappings (and the general difficulty in translating custom schemas consistently).

9 To reduce ambiguity, we use <fieldname>.

10 See, for example, DomainKeys Identified Mail (DKIM)

11 For a complete list of MIME types, please see the IANA Registry here: https://www.iana.org/assignments/media-types/media-types.xhtml

12 In 2005, the IETF put out RFC 4155 which included in its Appendix A a "default" MBOX format, though usage of the format predates this. Thus, MBOX files sourced from before 2005 are less likely to be fully-compliant with this specification.

13 https://docs.microsoft.com/en-us/openspecs/exchange_server_protocols/ms-oxmsg/b046868c-9fbf-41ae-9ffb-8de2bd4eec82 and https://docs.microsoft.com/en-us/openspecs/office_file_formats/ms-pst/141923d5-15ab-4ef1-a524-6dce75aae546

14 See another EA:BCC project, EA-PDF, for draft specification to address email records.

15 In some jurisdictions, judicial and legislative records are not covered by public records laws that apply to executive branches, confounding these situations substantially by introducing (or removing) requirements. PREPARE will focus primarily on executive branch records covered by public records laws.

16 https://www.statearchivists.org/viewdocument/dpcmm-assessment-insights-and-serp

17 Many states and territories address these potential shortfalls with email preservation, retention, and management guidance published outside of records schedules.

18 Of course, simply because a public official is part of an administrative process, that does not guarantee the existence of such records in their email, nor does it preclude the existence of other, unrelated records in their mailboxes. For a more complete discussion of how PREPARE recommends approaching these issues, see Section 3: Preservation which focuses on techniques to analyze the content of mailboxes.

19 Such as the use of natural-language processing or other artificial intelligence enabled tools.

20 https://github.com/StateArchivesOfNorthCarolina/tomes-eaxs

21 https://github.com/StateArchivesOfNorthCarolina/tomes-eaxs

22 Summary of Annual FOIA Reports for Fiscal Year 2021. Office of Information Policy, U.S. Department of Justice. Retrieved from: https://www.justice.gov/oip/reports-1