Emails are challenging records to preserve as they are valuable both as individual records and as aggregate collections. Planning how to structure and store email records can be essential to successful preservation of data.

PREPARE PROJECT

A Primer on Email Preservation Packages

One of the more challenging aspects of email preservation is the creation of preservation packages: structured groups of records that can be deposited in an Open Archival Information System (OAIS)-compliant repository or otherwise preserved for extended periods of time. Emails are challenging records to preserve as they are valuable both as individual records and as aggregate collections; therefore, planning how to structure and store email records (and similarly understanding legacy plans) can be essential to successful preservation of data.

The basic technology of email is generally quite simple: messages are sent from one email server to another, and organized and stored in the context of an account. Messages have required (and optional) metadata included in the message header, as well as a body of text and potentially other data. Additionally, one or more attachments (discrete files) can be included with email messages.

This simple specification for the interchange of electronic messages provides a great deal of flexibility which can present some challenges. Firstly, while there is a standard format for electronic messages[1], there is no standard implementation for tools that create, organize, and send such messages. Due to variations in the nature of the software that creates and sends email, the similarities between email records end with the creating program. Therefore, it is essential to understand the software that creates and receives email is an important part of its provenance.

Some messages are simply text. Others contain rich-text, common formatting modifications such as italics, bold, alternate fonts, colors, and other options. The formatting of information in the body of an email, and the text included, can be meaningful in itself. Furthermore, the organization of messages in an email account can typically be managed extensively by email software. Many email platforms allow users to create custom folders, subfolders, tags, and other "enhancements" to email that are outside of the standard specifications.

Simply put, not all email is created equally, and it's not stored equally either.

Simply put, not all email is created equally, and it's not stored equally either. Due to these variations, those who package email for long-term preservation must take the following details into consideration.

[1] https://datatracker.ietf.org/doc/html/rfc5322

Determine the Scope of Preservation: Content and Metadata

As with most archival appraisals, the email preservationist must first determine which elements of electronic mail have permanent value.

- Beyond messages and folders, what other data elements that are part of email platforms have long-term value?
- Are contact lists to be included?
- Are calendars to be included?
- Are other pieces of data to be included (such as task lists, email templates, signatures, etc.)?
- Are custom message headers or other metadata elements about emails valuable?

These questions must be answered, as the answers will provide guidance when making other choices below.

At the highest level, archive formats like PST or ZIP directories can contain more information than is contained in messages. For instance, a user's customizations to their own folders as well as certain application-specific metadata categories (such as read/unread) can be preserved more readily in formats used by their creating applications (such as PST for Outlook/Exchange) and may be lost upon conversion to other formats if care is not taken to identify and preserve this data.

Choose the Right Fit: File Formats

There are several file formats available for preserving email long-term. When choosing a format, several considerations apply:

- **Archive formats vs. message formats:** will messages be retained in structured archive formats like PST or flat MBOX files, or will they be retained as individual messages in a typical file structure, or both?
- **Handling attachments**: will the selected format preserve attachments with fidelity?
- **Open source vs. proprietary formats**: proprietary archive formats like PST and message formats like MSG will provide more capability to preserve software-specific customizations but may be burdened by copyright/licensing issues and/or interoperability issues in the future. Open source standard formats such as EML or MBOX are less able to support custom features but are much more likely to be readable into the future. There are many formats in which electronic message data can be readily preserved including XML and SQL-based databases.

Commonly-used formats:

The **Electronic Mail Format (EML)** is an open format designed to comply 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's 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.

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 a mailbox, so not multiple folders) in a single database file. MBOX files are useful since they collapse all messages into a single file, making it easier for applications to parse large swaths of data with few intermediate steps. One limitation with MBOX files is that historically there has been no "standard" syntax, and each application that creates MBOX files will do so slightly differently.[2] This can cause compatibility issues in the future if care is not taken to document what software is used to generate MBOX files, and which elements are captured in the file.

HTML is another format which has been used, especially in legacy contexts, to preserve email. Some legacy email applications used HTML extensively to create and render messages, essentially transforming email headers and body into visually-stylized HTML pages. Unfortunately, the drawback of these methods is that email software has significant variations in their IMF-to-HTML conversions, and these files can be very difficult to parse without comprehension of the original creating software. Difficulties with HTML may be avoided if metadata contained in message headers is specifically included in the file.

PST and MSG are commonly-found file formats in state and territorial governments given that a significant number of jurisdictions use Microsoft Exchange and Outlook as their email platforms. These formats provide a substantial amount of additional information about the user account and messages which can be valuable to capture. Fortunately, these standards are openly documented.[3]

Some might choose to use an extremely simple file format for preserving essential details only, such as the email headers and body. For such tasks, use of TXT, RTF, or PDF files can be useful; in particular, PDF has the ability to store embedded files (which could be email records) which could allow for useful packages of email sets in a single document. Similar warnings apply to these formats as does HTML -- care must be taken in order to ensure that the crucial properties of the form and content of email are not irretrievably lost.

In the future, new attempts at email archive formats such as EA-PDF[4] are promising efforts that may address some of the limitations of other formats.

[2] 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.

[3] 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

[4] https://www.pdfa.org/community/ea-pdf-lwg/

Structure Data, including Attachments

Another element to consider is the method used to preserve attachments, which will influence some of the other choices. The Multipurpose Internet Mail Extension (MIME) specifications[5] define the protocol for the transmission of non-textual data such as audio, video, images, and more.

The choice of file format (and software application) in which data is to be stored plays a major part in determining how attachments will be preserved. Some will preserve attachments within the message record, while others will separate them but maintain a contextual link.

One method commonly seen is to extract the attachments from the individual messages and store them elsewhere (this method is employed by many MBOX converter software). In this example, one such method might include mimicking the folder structure as presented in the email account (i.e. Inbox, Sent Items, Custom Folders, etc.) in a file directory with the individual attachments saved in these directories. Such as:

Email
  ↳ *Inbox*
    ↳ *2021-Messages*
    ↳ *2021-Attachments*
    ↳ *2022-Messages*
    ↳ *2022-Attachments*

One advantage of preserving attachments outside (e.g. not embedded) other email records is to maintain their accessibility by filesystems and other machine processes; for example, some digital repository software may be able to read the full-text of a PDF but not be able to extract data from records embedded within. Depending on the use-case, the archivist may wish to be able to appraise, analyze, and otherwise manage these records separately (especially if they have individual preservation issues).

Document Digital Provenance

Modern email software (including Outlook, Gmail, Thunderbird, and other applications) can create and expects to read email in particular formats. Some email applications, including those designed for archival processing and preservation, may not accept email that is not properly-formed.

Legacy email software (especially those used in the 1990s and 2000s) were much less standardized and powerful than today's applications, and many utilized complex and arcane methods to store data about messages, users, and other information.

Therefore, an essential part of email appraisal is the identification of the software used and methods for creation of email records. Conversely, archivists ingesting email into their repositories and creating archival packages must consider how users will use email software in the future, and how they might expect to understand the contextual links of extant email collections.

> ***An essential part of email appraisal is the identification of the software used and methods for creation of email records.***

[5] https://en.wikipedia.org/wiki/MIME

Conclusion

In sum, the major considerations for long-term email packaging are:

- Determining the scope of essential information to be preserved
- Matching format to scope and functions
- Structuring data to preserve context
- Documenting digital provenance

Additionally, these considerations merely address the structure and format of a set of email; typical preservation actions such as the creation/verification of fixity, migration, and so on must still occur beyond the decisions made here, and thus those processes should also broadly inform these choices.

There is no best approach, only best fit based on requirements, since a myriad of factors (IT centralization, software adopting, email administration) can change the context of email's creation, storage, and use.

Email is challenging for state and territorial governments to manage due to the wide range of possibilities involved in the manner and tools used to create and ultimately store email. There is no *best* approach, only *best fit based on requirements,* since a myriad of factors (IT centralization, software adopting, email administration) can change the context of email's creation, storage, and use. Therefore, studying the underlying considerations can be the best approach to building an email preservation framework from the ground up.

About CoSA

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.

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 CoSA is providing to state and territorial archives as part of the State Electronic Records Initiative (SERI).

August 2022