Recent developments on email preservation: towards the ultimate solution?

Stefano Allegrezza [1]

[1] *University of Bologna, Italy*

Abstract

This paper aims to provide an overview of the email preservation problem that is still unresolved. Starting from a review of the main preservation challenges (such as the huge amount of emails sent and received, the issue of attachments in a variety of formats, the issue of links to external resources, the issues relating to privacy and digital legacy, etc.), it deals with the email preservation strategies that have been proposed in the last thirty years (from printing to paper, to normalization to XML based formats, to the use of PDF and PDF/A and its various profiles) none of which are the perfect solution. Finally, the most interesting recent developments on email preservation are highlighted. In particular, great attention is given to the EA-PDF format that promises to become the ultimate solution to the problem of email archiving and preservation.

Keywords

digital preservation, email archiving, email preservation, PDF, PDF/A, EA–PDF

1. Introduction

Email was invented by Ray Tomlinson in 1971, when he sent himself a elementary test message in 1971 – something like 'QWERTYUIOP' – and in 50 years it has grown to become one of the most widely used means of communication for both personal and business interactions. It began to replace paper–based personal and business correspondence in people's work and personal life in the late 1980s and early 1990s. Nowadays email is essential to most people's work and personal lives, and it is a quick and easy way to transmit private and professional messages and thoughts. The messages that we send and receive leave behind an evidence–rich trail of actions, thoughts, and communications [1].

2. Why it is important to archive and preserve emails

Email documents the personal and public stories of the day. From «family gossip to friendly chatter to institutional business decisions to government actions, all are now frequently documented in email accounts across the globe» [1]. For many organizations, email accounts may be one of the most important sources of documentary evidence of activities and transactions, and therefore have enduring archival value [2]. While it's true that most emails have little long-term storage value, some absolutely do and should be preserved. Therefore, email is a source for documenting the history and reconstructing it in the future, and can give valuable documentation of significant transactions and decision–making data [1]. In recent years email was highlighted in many cases as a source of information for personal, political, business, and academic issues in the news [1]. As a result, maintaining access to emails may be critical to ensure that those in positions of power are held

IRCDL, February 24-25, 2022, Padua, Italy
EMAIL:
ORCID:

accountable. It's also a significant historical and social record, with researchers demonstrating that they can build portraits of social and corporate networks using email interactions [2].

Despite the fact that most archivists recognize the historical or cultural value of email, just a few institutions have made substantial progress in preserving it. A tiny number of repositories have taken on the task of long–term preservation, and an even fewer number have built policies, implementation methods, procedures, tools, and services to accomplish so.

3. Email preservation challenges

Email is particularly difficult to preserve, putting future access to this vast resource at risk. In fact, archiving and preserving email isn't easy or obvious. Email is a complex data type consisting of multiple conversations involving many different people [3]. It can include attachments (which may be of any other data type or file format), it can be large in size and it can be challenging to capture it effectively. Email messages appear transitory and ephemeral when compared to other types of documents, such as writings, presentation, photographs, etc. which can be viewed with little technical aid. Commonly, solutions are vendor–specific and email clients are required: not an ideal solution for static records. Furthermore, email is characterized by ubiquity but also ephemerality, both of which can cause problems for long–term preservation.

One of the main critical issues is the huge amount of emails that are sent and received everyday. Email volume has grown exponentially and hundreds of billions of emails are exchanged around the world every day [4]. As a consequence, the size of the email archive an institution needs to manage easily reaches the order of tens of GB and contains tens of thousands of emails. This poses serious problems for both the management and preservation of emails.

Another big problem is the preservation of attachments. From a technical point of view an email consists of three parts, all of which need to be preserved: the header that is basically the equivalent of the letterhead and contains the sender and recipient information, the creation date and some optional information such as the subject in the form of metadata; the body, i.e. the actual email content, that is displayed differently depending on the user–defined settings in the email software; and, optionally, one or more attachments, that are often text documents, spreadsheets, presentations, technical drawings, images, scanned documents, etc., possibly combined in a ZIP file. Almost any file type can be included as an attachment, so exotic file formats or executable programs or scripts can also be included. The real problem in email preservation is not the preservation of the header or the body (that are important, of course), but the preservation of attachments. Email attachments may need to be preserved along with the body of the email to which they were attached. While preserving the body is relatively easy, preserving the attachments may be very difficult, for instance due to the use of obsolete or proprietary and non-standard file formats. Preserving an email archive may therefore require the preservation of potentially large amounts of other digital files in a variety of formats. In addition, both body and its attachments must be preserved in such a way as to maintain the relationships between them.

Another problem is the presence of external links. Emails can be incredibly complex, containing references to external content which is required to understand the message itself. For example, in an email there might be an image that is not embedded in the email itself but is simply retrieved from the web as the email is viewed. If in the future that image is deleted, renamed, or moved to another folder, it can no longer be displayed in the email. Links tend to last very little and after a short time they are often broken; therefore in order to preserve the visual appearance of an email, links to external content need to be "resolved" and the external content must be included in the email itself.

Obviously there are privacy issue as well. The body of an email may contain personal or sensitive information. Email attachments may also contain such information or other personal data. There may also be copyright complications, as the email author is not the email receiver. Reuse and retention may be affected by varying legal frameworks in different jurisdictions [5]. Email accounts can hold significant volumes of personal information and sensitive data because of the nature of the organization (such as banks, medical offices, or certain types of government agencies) or position of the person (for example, human resources personnel, executives, or union officials) [3]. Private

individuals donating email accounts to archives may have used their email to correspond with any number of other private individuals who are unaware that their communications are being preserved.

Another important aspect to consider is the increasingly worrying 'digital legacy' issue. When the creator dies, it is often very difficult to have access to the email archive of the deceased, because usually nobody share with others the login credentials to access the mailbox. In addition, email service providers usually keep emails only for a specified period of time – usually ranging from a few months to one or two years at most – since the last time the mailbox was accessed. So, if you wait too long before you try to recover the email archive of a deceased person (or someone who has not accessed his account for a few months), you will find the mailbox completely empty.

Another issue is the lack of policies. Most organizations lack clear policy and guidance for email users on how to manage and preserve email, meaning that everyone can take a different approach.

Finally, we must consider the lack of a specific preservation format. Since there is no 'native' preservation format, any effort to preserve email requires some kind of transformation when extracting it from its originating system. The format used by the software to access and manage email is not standardized. This means that email formats can be dissimilar and different and there is no consensus on which preservation format to use.

What we've just seen are just a few of the issues we need to face when we want to implement an email archiving and preservation strategy and that make this task quite difficult.

4. Email preservation challenges

Over the past thirty years, several solutions have been proposed for archiving and preserving email. One of the solutions still widely used consists in printing the email to paper. Printing emails on paper may appear a simple and straightforward solution, but it is an anachronistic and inefficient solution, without considering the ecological consequences as well. Attachments and context in connection to other messages (such as threads) may be lost, as well as the contents of email headers. As practical experience and digital preservation have advanced, this 'print and file' approach to email archiving is now recognized as destructive, resulting in a loss of contextual information such as metadata, changes to the look and feel of email and the associated user experience, and dissociates messages from their attachments.

Another widely used strategy is based on the conversion of emails to a system–independent format using one of the many specialist tools available. The majority of these neutral formats are based on XML. The context/relationship of an email with other emails will also be recorded. But even XML–based formats have their limitations, such as in the preservation of the visual aspect. So, it is recommended that careful research and analysis is carried out before extracting emails or migrating them to a new format.

Other strategies rely on the adoption of some preservation formats, such as the EML format for single emails and the MBOX format for aggregations of emails (typically the contents of a folder). For example, Library and Archives Canada (LAC) and the US National Archives and Records Administration (NARA) recommend this approach. However, both of them are not specific preservation formats. In addition, due to the large volume of emails that can be contained in a single folder, MBOX files can often become excessively large and unwieldy; it also means that, in some cases, the corruption of a single message can prevent the entire file from loading or opening. MBOX as a preservation format also suffers from the fact that actually it is not a single format but a family of formats with four variants (MBOXO, MBOXRD, MBOXCL and MBOXCL2) that are not entirely compatible with one another. Other formats, such as MSG and PST, are developed by Microsoft and although they are widely used and Microsoft has published their specifications, they are proprietary and subject to frequent changes. Keeping email for archival purposes using often unfamiliar and/or proprietary formats is a daunting task. Anyway, to date no single perfect format has been proposed for the preservation and future use of email. Decisions made on file formats should be dependent on the features and functionality to be preserved and the future use cases to be supported.

Another widely used strategy consists in 'printing' the email and its attachments to a PDF file. Many people think that converting emails and their attachments to PDF will ensure they remain readable for long periods of time. Unfortunately, PDF is not a preservation format: for example, it

does not require the incorporation of fonts, so it cannot guarantee the preservation of the visual representation of the body of the email and its attachments. In addition, all the information that guarantees the provenance and authenticity (typically contained in the header) is not captured.

To solve this problem, system–independent archiving of emails and attachments in PDF/A–1 format (ISO 29500–1) is recommended [6]. This format has long been established for general archival purposes and is now recognized as the standard for long-term preservation for any kind of document that can be printed. Compared to PDF, PDF/A has several features that make it suitable for long-term storage. For example it requires the embedding of fonts (for text) or ICC profiles (for images) in order to ensure the reproducibility of emails over the years. In other words, PDF/A ensures that an email and its attachment will be displayed exactly the same way on any system; in addition it forbids any dynamic content. Furthermore – and this is the most important argument of all – the ISO standard is designed for long–term archiving, guaranteeing that emails saved in the PDF/A format will remain reproducible and readable for decades to come [6]. Obviously this assumes the availability of a PDF/A viewer for as long as that email is to be kept, but the format is so widespread that we can consider it assured.

With the release of the PDF/A–2 format (ISO 29500–2) in 2011 it became possible to convert the body of an email in PDF/A–2 and the attachments in PDF/A and then incorporate them into the PDF/A–2, thus obtaining a fully preservable object. The PDF/A–3 format (ISO 29500–3), released in 2012, allows to incorporate in the principal PDF/A–2 not only the PDF/A file obtained from the conversion of attachments, but also the attachments themselves in their native format, thus obtaining an object that is suitable for reuse of original file formats and at the same time suitable for preservation needs.

Recently, the PDF/A–4f conformance level of the PDF/A–4 (ISO 29500–4) released in 2020 has become available as the successor to PDF/A–3. The new version of the PDF/A format allows embedding any file, like its predecessor PDF/A-3 [7]. On this basis, at least the issue of identifying a format for archiving email can be answered quite satisfactorily, as most email systems offer an export function to PDF or PDF/A. Unfortunately, however, this approach often falls short, because usually only the email body is taken into account and not the header or any attachments. If emails are to be archived in PDF in their entirety, the header should be saved as XMP metadata in the PDF file. This can then be used as the basis for a targeted search for emails. The email body is ideally converted on the basis of the body type (plain ASCII, formatted text, HTML) that most comprehensively reflects the contents. Links or referenced images in HTML must then also be integrated. The greatest flexibility in the use of archived emails is available if the original email file in EML or MSG format and the attachments are also embedded in the PDF, which is possible with PDF/A–3 or PDF/A–4f [7]. Experience has shown that emails archived as PDF/A (in any of the available versions) are almost always larger than the original files, due to the fact that the PDF/A standard requires the embedding of fonts or ICC profiles for colors in order to ensure the reproducibility of emails over the years.

In short, there is still a long way to go towards an effective solution, but some interesting solutions are beginning to be proposed.

5. Recent developments on email preservation

In the last few years the digital preservation community has been very focusing on the topic of email preservation and several projects have come out, some of which are worth mentioning.

One of the most interesting project was led in the period 2016–2018 by the Task Force on Technical Approaches for Email Archives, funded by The Andrew W. Mellon Foundation and the Digital Preservation Coalition [8]. In august 2018 the Task Force released a report of their findings: “The Future of Email Archives” [1]. Published by the Council on Library and Information Resources (CLIR), this report is intended for the archival community, digital preservation professionals, technologists and software developers, commercial vendors, historians and scholars, institutional administrators, and funding agencies and foundations. It provides a detailed analysis of the technical challenges to preserving email and proposes a working agenda for the community to improve and refine the technical framework for email archiving, including developing interoperable toolkits to fill in the missing gaps. The task force proposes a series of short– and long–term actions for community

development and advocacy, as well as for tool support, testing, and development. One of the gaps identified and a goal of many of the other email archiving projects is to identify a format or a list of formats that are appropriate for use in archiving email for long–term preservation. One strong contender identified in the Email Task Force report was PDF, and specifically, the PDF/A subset.

Another interesting project is RATOM (Review, Appraisal, and Triage of Mail) at the University of Carolina [9]. RATOM is developing software to assist archives and other collecting organizations with email analysis, selection, and appraisal tasks. In fact, despite progress on various technologies to support data management and digital preservation, relatively little progress on software support for the core activities of selection and appraisal has been made. Anyway, as selection/appraisal decisions are based on various patterns, software can assist the process if patterns can be identified algorithmically. The project extends the email processing capabilities currently featured in the TOMES software and BitCurator environment, developing additional modules for these tools along with specific standalone software to support more advanced workflows. These include identifying and reporting on entities within emails and email attachments using a scalable NLP library; identifying materials requiring review due to the presence of potentially sensitive information; and developing software modules to assist with preparation of materials for release and public access.

A project that is worth mentioning is COPTR (Community Owned digital Preservation Tool Registry) [10]. COPTR is a wiki based registry of digital preservation tools to assist email preservation. It's main aim is to help practitioners to discover preservation tools that will help them to tackle particular preservation challenges. It can be browsed and searched directly by practitioners, or queried by other systems via an API. Each tool in COPTR is categorized by the lifecycle stage it falls within, the (sub) function it performs and the content type that it operates on. It may also input and/or output specific file formats. As of October 2021, it describes 541 tools; of those, 12 are tools for working with email, such as Aid4Mail, DArcMail, Emailchemy, MailStore Home.

Another interesting project is ePADD, led by Stanford University's Special Collections & University Archives to develop a free and open source software designed to support the appraisal, processing, preservation, discovery, and delivery of historical email archives [11]. ePADD incorporates techniques from computer science and computational linguistics – including machine learning, natural language processing, and named entity recognition – to address challenges that collection donors, archivists, and researchers routinely face in donating, administering, preserving, or accessing and searching email collections of historical and cultural value. This includes screening email for confidential, restricted, or legally–protected information, preparing email for preservation, and making the resulting files (which incorporates preservation actions taken by the repository) discoverable and accessible to researchers. Initial work on ePADD began in 2010 and received funding from the National Historical Publications & Records Commission from 2012–2015 to develop the first full version of the software package. From 2015–2018, the project received funding from the Institute of Museum and Library Services to develop a further six versions of ePADD. To continue development on the software, from 2020–2021 the ePADD project received funding from the Andrew W. Mellon Foundation. In 2021, the ePADD project received funding from the Email Archiving: Building Capacity and Community grant program, administered by the University of Illinois at Urbana–Champaign to integrate preservation functionality into ePADD.

Currently the most interesting project in the field of email preservation is the proposal of a PDF specification for email preservation, called EA–PDF (Email Archiving PDF). The project is a collaboration between the University of Illinois, the U.S. National Archives and Records Administration (NARA), the Library of Congress, the PDF Association and others [12], and was funded by Andrew W. Mellon Foundation project in 2019. Its goal was to investigate how email messages and their identified essential characteristics and functionality should be converted into PDF containers that can be considered – in the context of captured information – authentic and complete email records [13]. This led to the publication of a final report, "A Specification for Using PDF to Package and Represent Email" [14] which introduces the EA–PDF concept and establishes high–level functional requirements for using PDF technology as a model for packaging email for long–term preservation purposes, both for individual emails and for complete folders or entire mailboxes. These requirements detail desirable functionality reflecting considerable input from stakeholders in digital preservation, government, education and industry communities and define the significant characteristics of email required to meet the needs of the email archiving community. Conceptually,

EA–PDF is no more complex than the underlying source email, but represents that complexity in a formally–defined manner, within the structures of the PDF container (ISO 32000), while other email formats, such as EML and MBOX are less well–defined family of formats defined more by client implementations than by authoritative specifications. In addition, PDF's ubiquity and acceptance, its reliability and interoperability, its rich capabilities and open, well–documented specification is already supported by a global ecosystem of developers. At the end of 2021 the University of Illinois funded the PDF Association with a 24-month project to develop a detailed technical specification defining the interoperable use of PDF (ISO 32000) as an archival format for email. A working group, called "Email Archiving Liaison Working Group" (EA-LWG) was formed as a forum for focusing industry engagement on the specification in partnership with experts from archives and libraries that are interested in contributing to the development of the EA-PDF specification. In a separate effort the University of Illinois will also fund development of an open-source proof-of-concept implementation of the specification.

In conclusion, EA-PDF seems to be the real definitive solution to the problem of long term email preservation.

6. Final tips and conclusions

There are a few suggestions that can be given to facilitate email preservation.

From the point of view of institutions undertaking email preservation projects, they should define policies, choose the appropriate tools, and implementing them according to local environmental factors and available resources [5]. First of all, institutions should put in place email policies and guidelines. Email policies should state the institution's commitment to email preservation as well as the steps that will be taken to put it into action. Email categories that are significant for administration, institutional memory, and cultural value must be defined by policies. Obviously, creators working in the institution should adhere to the policies.

The second step after defining policies, is choosing the tools to use for appraisal/capture; for the conversion from the native email format to a preservation format; for arrangement (if needed) and description and, finally, for storage in the preservation system and access. As previously said, there are numerous options, and the Email Archives Task Force Report [1] offers a useful list of tool and descriptions that were current at the time of writing.

The last step is the implementations of policies. In practice, institutions will need to establish a protocol for email processing. Emails may come from a variety of systems or accounts, thus translating them to a standard format, such as MBOX or EA–PDF, the new standard being proposed, will be a critical first step. A more or less linear workflow can be constructed from there.

From the point of view of creators, managing and preserving email can be difficult for many of them but following some elementary rules can guarantee that email records of lasting historical importance are managed effectively. An important rule to keep in mind is to maintain personal and professional email separate; in other words, it is preferable not to use your professional or work account to send and receive personal emails, and vice versa. Another rule is to organize your email archive setting up folders for filing important emails (for example, all the emails sent and received with a particular correspondent or all the emails relating a particular project or activity) – instead of maintaining all the emails in the Inbox and Sent folders. And, obviously, they should do this on a regular basis, even if this a time–consuming activity.

Email preservation is not simple task. Addressing the challenges will require commitment and engagement of a wide variety of stakeholders, from the creators to the records managers, librarians, archivists, curators, IT professionals, organizational leaders or other information professionals. If we all collaborate we can success in preserving email for its legal, administrative, cultural or historical value. In short, there is still a long way to go towards an effective solution, but some interesting solutions are beginning to be proposed and perhaps the EA-PDF Project could lead to the ultimate solution.

7. References

[1] CLIR - Council of Library and Information Resources, “The future of email archives, A Report from the Task Force on Technical Approaches for Email Archives”, August 2018, https://www.clir.org/pubs/reports/pub175.
[2] InterPARES3 Project, “General Study 05 – Keeping and Preserving E–mail”, June 2009, http://www.interpares.org/ip3/display_file.cfm ?doc=ip3_italy_gs05a_ final_report.pdf.
[3] DPC - Digital Preservation Coalition, ”Preserving Email. Digital Preservation Topical Note 7”, 2018, https://www.dpconline.org/docs/knowledge–base/1868–dp–note–7–preserving–email/file.
[4] The Radicati group, inc., “Email Statistics Report”, 2015–2019, https://www.radicati.com/wp/wp–content/uploads/2015/02/Email–Statistics–Report–2015–2019–Executive–Summary.pdf.
[5] Prom, Chris, “Preserving Email (2nd Ed)”, DPC Technology Watch Report 19–01 May 2019, https://www.dpconline.org/docs/technology–watch–reports/2159–twr19–01/file.
[6] PDF Association, “Email archiving with PDF/A”, January 14, 2015, https://www.pdfa.org/email–archiving–with–pdfa.
[7] Von Seggern, Dietrich, “Emails for eternity”, July 14, 2021, https://www.pdfa.org/emails–for–eternity.
[8] Prom, Chris, “The future of past email is PDF”, February 28, 2020, https://www.pdfa.org/the–future–of–past–email–is–pdf.
[9] RATOM (Review, Appraisal, and Triage of Mail) Project, https://ratom.web.unc.edu.
[10] COPTR (Community Owned digital Preservation Tool Registry) Project, https://coptr.digipres.org/index.php/Main_Page.
[11] ePADD Project, https://library.stanford.edu/projects/epadd.
[12] Duff, Johnson, “Archiving email into PDF containers: A Mellon Foundation project,” July 10, 2019, https://www.pdfa.org/archiving–email–into–pdf–containers–a–mellon–foundation–project
[13] PDF Association, “Packaging email archives using PDF”, February 11, 2021, https://www.pdfa.org/packaging–email–archives–using–pdf.
[14] EA–PDF Working Group, “A specification for using PDF to package and represent email”, University of Illinois at Urbana–Champaign, 2021, https://www.ideals.illinois.edu/handle/2142/109251.