**ISSN: 2048-7916**

**DOI: http://doi.org/10.7207/twr19-01**



First published in Great Britain in 2011 by the Digital Preservation Coalition. Second edition 2019.

## Foreword

The Digital Preservation Coalition (DPC) is an advocate and catalyst for digital preservation, ensuring our members can deliver resilient long-term access to digital content and services. It is a not-for-profit membership organization whose primary objective is to raise awareness of the importance of the preservation of digital material and the attendant strategic, cultural and technological issues. It supports its members through knowledge exchange, capacity building, assurance, advocacy and partnership. The DPC's vision is to make our digital memory accessible tomorrow. The *DPC Technology Watch Reports* identify, delineate, monitor and address topics that have a major bearing on ensuring our collected digital memory will be available tomorrow.

They provide an advanced introduction in order to support those charged with ensuring a robust digital memory, and they are of general interest to a wide and international audience with interests in computing, information management, collections management and technology. The reports are commissioned after consultation among DPC members about shared priorities and challenges; they are commissioned from experts; and they are thoroughly scrutinized by peers before being released. The authors are asked to provide reports that are informed, current, concise and balanced; that lower the barriers to participation in digital preservation; and that are of wide utility. The reports are a distinctive and lasting contribution to the dissemination of good practice in digital preservation.

This report was written by Chris Prom, Associate Dean for Digital Strategies, University Library at the University of Illinois at Urbana-Champaign. The report is published by the DPC with the support of the Research & Practice Subcommittee who provide editorial oversight.

## Acknowledgements

The author would like to thank the following people, who have either inspired or directly helped with this report: Kate Murray (Library of Congress), Courtney Pierre Joseph (Lake Forest College), Grace Moran (University of Illinois), Sara Day Thomson, William Kilbride, members of the Task Force on Technical Approaches for Email Archives, and the anonymous reviewers whose comments helped me shape the report's content and recommendations.

## Contents

## 1. Abstract

This report reviews the current state of email preservation and offers recommendations for information professionals (such as organizational leaders, records managers, IT professionals, librarians, archivists and curators) who seek to preserve email for its cultural, legal or administrative value. It also provides guidance to private individuals who may wish to preserve their email correspondence and to deposit it in a cultural heritage institution. Whatever choices that people and repositories make, this report describes the key policies, implementation strategies, procedures, tools, and services that can be drawn upon when developing an email preservation programme. Ultimately, it presents a hopeful message: by implementing appropriate technical standards, capture methods, and processing tools, every archivist, curator, records manager, or other information professional can take practical steps to preserve email for its legal, administrative, or historical value.

## 2. Executive Summary

Since the first edition of this report was published in 2011, email preservation has become a common, though not yet routine, part of digital preservation and archives work. In some ways, this is not surprising, since the core issues have changed little over eight years. As essential as email is to conducting business, many organizations treat non-current messages as more of a liability than an asset. And as important as email is to facilitating personal communication, many individuals find it as much a nuisance as a blessing.

The messages that we send and receive daily leave behind an information and evidence-rich trail of our activities. Information professionals such as archivists and records managers can choose from a range of tools, both open source and commercial, to complete the basic tasks of appraisal, disposal or acquisition, arrangement, description, migration, storage, and discovery. These actions make email messages and accounts accessible resources, able to be interrogated and used for administrative, legal, and historical purposes.

This report describes specific ways in which institutions have made significant progress in implementing these tools, as well as more advanced tools, including some focusing on natural language processing, machine learning, and predictive coding. As engaging as these new options might seem, they complement rather than replace traditional approaches, and most archives will be best served by starting with good lifecycle management and preservation services using relatively common preservation tools. And there is more good news: since email messages and attachments can be converted to relatively stable, preservation ready formats, they can be managed and preserved within an institution's existing repository infrastructure, everything from a replicated file-based system to the most complex, cloud-based preservation service. While archivists must make many decisions when developing an email preservation programme, this report described options and offers implementation advice suited to a range of institutional circumstances.

## 3. Introduction

Since the first edition of this report was published, email preservation has become a more common, though not yet routine, part of digital preservation and archival work. In some ways, this is not surprising, since the thorny non-technical issues noted in 2011 still apply today. As essential as email is to conducting business, many organizations treat old messages as more of a liability than an asset. And as critical as email is to facilitating personal communication, many people see its constant drip more as a nuisance than a blessing. Even if people dutifully file messages away, they find it easy to forget about those they have read or sent, at least until they need to dredge up an important report or a reminder of some past decision.

In our information-saturated lives, email seems to be but one of many communication choices, jostling for attention alongside text messages, chat, calendar notifications, voicemail, and social media postings. Now email also exists alongside workflow and collaboration systems, such as Slack[1] and Trello[2], which replicate and extend much of email's functionality. In fact, all the systems just described leave traces in people's email accounts, as they send notifications, reminders, and updates. This implies that while a mature digital preservation program focusing on communications should take account of more than just email, those looking to preserve digital communications should start with it, since email is still a lynchpin technology.

[1] Slack: https://slack.com/
[2] Trello: https://trello.com/en

Accordingly, this report reviews the current state of email preservation and offers recommendations for organizational leaders, records managers, IT professionals, librarians, archivists and curators (a group hereafter referred to as information professionals) who seek to preserve email for its cultural, legal or administrative value. It also provides guidance to private individuals who may wish to preserve their email correspondence and to deposit it in a cultural heritage institution.

## 3.1. Importance of Email Preservation

In 2014, tech writer Alexis Madrigal noted that although the then-current wave of Silicon Valley wunderkinds was predicting email's demise as a 'counter-productivity tool', its usage was likely to persist into the indefinite future (Madrigal, 2014). Madrigal was more right than he could have known. Not only does it persist, it thrives. A group of economists recently calculated the implicit value that people accord to seemingly free Internet technologies such as maps, social media, and instant messaging. Based on their survey results, email was the second most indispensable service, after search engines (Brynjolfsson, Eggers, and Gannamaneni, 2018, p. 30). Respondents indicated that they would require, on median, a payment of nearly 8,500USD to give up it for just one year (Garcia, 2018)!

Email is essential to most people's work and personal lives, and the messages that we send and receive leave behind an evidence-rich trail of actions, thoughts, and communications. In 2011, the first edition of this report pointed out that investigative journalists, muckrakers, and agents provocateurs love email messages (Prom, 2011, p.5). Subsequent news reports indicate that email is still regarded as a valuable – and often compromising – type of documentation (Task Force on Technical Approaches for Email Archives, 2018a, pp. 5–7). It is both a story keeper and a storyteller, containing many embedded and entangled narratives (Chapin and Attfield, 2018). For instance, it served as a core source for the LIBOR investigators (the *Telegraph*, 2013; the *Guardian*, n.d.), and email continues to be routinely cited in news exposés (Bernstein, 2017; Stripling, 2018). Some authors even ascribe the results of the last United States Presidential Election to missing emails – or at least to how those missing emails were perceived – in what might be seen as history's most consequential records management snafu (Helderman and Hamburger, 2016). And it continues to haunt the political landscape; at the time of writing the US President's daughter was caught in a minor email imbroglio (Stewart, 2018).

While most archivists acknowledge the importance of email, a relatively small number of institutions have made significant progress in preserving it for historical purposes. A subset of repositories has embraced the

responsibility to preserve it for the long term; an even smaller number have developed policies, implementation strategies, procedures, tools and services that systematically do so. And even some of our most well-respected records management and archives programs have suffered severe criticism for their inability to better control this most troublesome resource (Bearman, 2017). But this is not for lack of technical knowledge or advice.

This report describes the specific ways institutions can make significant progress in implementing common email preservation tools as well as more advanced software, including some focusing on natural language processing, machine learning, and predictive coding. These engaging new options complement traditional approaches, and most archives will be best served by starting with good lifecycle management and preservation services using relatively common and stable preservation tools. And there is more good news: since email messages and attachments can be converted to relatively stable preservation formats, they can be managed and preserved within whatever existing infrastructure a repository has developed, everything from a replicated file-based system to the most complex, cloud-based preservation service.

Whatever choices an archive makes, this report describes the key policies, implementation strategies, procedures, tools, and services that can be drawn upon. Ultimately, it presents a hopeful message: by implementing appropriate technical standards, capture methods and processing tools, every archivist, curator, records manager, or other information professional can take practical steps to preserve email for its legal, administrative, or historical value, building on work completed over the past 30 years.

## 3.2. Overview of Work to Date

Many authors have provided informal advice to those seeking to manage personal email more effectively (Schmitz Fuhrig, 2011; Guy, 2011; Ashenfelder, 2011a). However, until the last five years, relatively few formal reports or articles discussed the topic of email preservation. A few individuals had previously called for libraries, archives and museums to preserve email correspondence for its cultural value (Hyry and Onuf, 1997; Enneking, 1998; Marshall, 2007; Cox, 2008). And the first edition of this report cited a few essential publications providing specific implementation advice (Paquet, 2000; Mackenzie, 2002; Schmitz Fuhrig and Adgent, 2008; Goethals and Gogel, 2010).

Two early works are still relevant: David Bearman's 1994 article 'Managing Electronic Mail' and Maureen Pennock's 2006 entry in the Digital Curation Centre's *Digital Curation Manual*, 'Curating E-Mails: A Life-cycle Approach to the Management and Preservation of E-mail Messages'. Each of these looks at a range of cultural, legal, ethical, professional and technical considerations that must be addressed if an organization wishes to identify email of permanent value, preserve it in an authentic form, and render it for future use.

Bearman argued persuasively that because email is governed by so few conventions, those wishing to preserve it must use a holistic and systematic method that preserves its value as evidence regarding a particular decision, function, or activity. Writing from the perspective of an archivist but with an eye toward business process analysis, he proposed that institutions pursue four strategies to implement this goal:

- educate users about email system operations;
- analyse the organization's objectives, structures and workflows, in order to identify functions or activities that must or should be documented;
- design systems to capture 'record' messages that document these functions or activities; and
- develop and deploy standards that support the long-term preservation of email records for their evidential value.

His approach has deeply influenced attempts to capture and preserve email, and some of his specific recommendations seem more feasible now than they did even five years ago. For instance, he recommended the development of automated tools to filter and capture 'record' emails into electronic records management

systems (ERMS), which might have an external store for both structured data and for semistructured records like email. The development of email journaling systems and application programming interfaces (APIs), discussed in more detail in Section 6 of this report, shows that this approach holds continued relevance.

Maureen Pennock focuses on a range of policy, design and implementation choices that an institution will face in attempting to preserve email. She reviews the legal and regulatory environment; articulates roles and responsibilities for those who use email, those who manage technical systems, those who wish to curate email archives and those who wish to use preserved messages; lists policy and technical options; and makes some practical recommendations. Specifically, she encourages institutions to educate users about their role in email preservation, to implement a method to capture email messages with long-term value, and to store them in a trusted repository, if possible, using a standardized XML format (Pennock, 2006, pp.31–33).

Since 2011, the field has developed a more systematic research agenda, theory, and practice of email preservation. A relatively complete bibliography of email-related writings can be found on the Task Force on Technical Approaches for Email Archives website. Several works are worth particular note.[3] Many of them provide practical methods that archivists and collection managers can use to acquire and manage emailbased collections. The Library of Congress has placed a particular emphasis on such advice (Lazorchak, 2013a, 2013b; Murray and Engle, 2015). The Society of American Archivists provides similar help, focusing on the ePADD software (Schneider, 2016; Coburn, n.d.), and the Carcanet Email Preservation project has also provided useful advice in its final report (Baker, 2014; 2015).

The most notable advance since 2011 has been in email-related research and policy work, with particular emphasis on how the archival value of email is appraised (see, for example, Cocciolo, 2016). In 2016, the United States National Archives and Records Administration (NARA) introduced its 'Capstone' email policy, which provides US government agencies with the option to appraise and preserve email based on the role of the account owner (Ferriero, 2016; National Archives and Records Administration, 2016). In developing the Capstone guidance, NARA acknowledged that it can be difficult for many agencies to manage email under existing records schedules, where records are treated by function or activity, not type or format. In contrast, the Capstone policy encourages agencies to appraise and capture entire accounts, or portions thereof, for agency leaders or decision makers, instead of conducting message-by-message appraisal based on message's content. NARA's policy shift inspired similar approaches in US government, academic, and business archives.

Current policy in the UK, by contrast, suggests that email messages should be assessed for their record value. The small percentage of emails with archival value should be filed alongside other records, in accordance with an organization's records schedules, potentially with the application of auto deletion procedures (UK National Archives, n.d.). This approach has received criticism, both in the UK and in the United States, since it has seemingly led to the preservation of very little email (United States Department of State, Office of the Inspector General, 2016; Lappin, 2015, 2018). However, as Case Study 7.1 demonstrates, records-based assessment is a viable option when exercised judiciously and in the light of General Data Protection Regulation (GDPR) requirements.

Other work, of a more technical nature, has assessed the ways in which email authenticity is understood by users (so that archivists can better assure it), as well as how advanced computational methods can be used to tag, sort, and weed email accounts of sensitive or non-record materials (Bunn *et al*., 2015; Cormack and Grossman, 2014; 2017). Of particular note is work by Harvard University (funded by the Andrew W. Mellon Foundation) to better situate email processing tools and workflows within the context of the email and records lifecycle (Harvard University, 2016; Simpson, 2016). Many of the conclusions from this work were

[3] Bibliography of email-related writings in the 'Task Force on Technical Approaches for Email Archives': http://www.emailarchivestaskforce.org/documents/bibliography/

incorporated into, and expanded upon, in the work of the Task Force on Technical Approaches for Email Archives (2018b), which recommended future development and technical work for the community.

This work – the 2nd edition of *Preserving Email* —complements the work of that Task Force, which was sponsored by the Digital Preservation Coalition and the Andrew W. Mellon Foundation. Rather than reporting new research or setting the forward development agenda, this *Technology Watch Report* attempts to summarize issues, assess standards and technologies, and recommend practical methods that can be used to acquire, process, preserve, and provide access to email correspondence – no matter the complexity or size of the institutional environment. It provides a particular focus on steps that institutions and individuals can take in order to implement 'good enough' preservation, using current technologies to capture email, migrate it to system-neutral formats, and store it in an accessible digital repository. As such, the *Report* draws on efforts from the records management, archives, digital preservation and computer science communities, to outline options that individuals and institutions can use to preserve this most important genre of records. Of particular use will be the targeted advice, along with case studies highlighting specific ways that three institutions are assessing, surveying, classifying, and managing email to identify and preserve those of archival value.

# 4. Issues

Email systems are simply communication utilities that can be used to send any kind of content, and email is one of the most successful and heavily used of all Internet technologies. These facts give rise to four challenges that confront anyone seeking to preserve email for its long-term historical value:

- storage formats, storage systems, institutional policy and personal management practices are extremely malleable both across, and within, institutions;
- most of the storage and retrieval costs for email fall upon people or institutions that do not have an interest in long-term preservation, such as line IT staff or companies providing free personal email accounts;
- legal requirements, resource constraints and lax personal information management practices lead many institutions and people to passively neglect or to actively delete email;
- from the end user's point of view, email is free, ubiquitous, and commonplace. Few people prioritize email management, much less its preservation.

Each institution must develop its own rationale for email preservation in light of its needs, profiles, mandates and policies, but the chosen approach must be based on an understanding of the technical, legal, organizational and personal factors that affect the preservation of this ubiquitous, ephemeral and evidencefilled documentary genre. A number of issues, rather than any one factor, need to be considered when preserving email, particularly if an organization's goal is to do so on a wide scale.

## 4.1. Legal and Policy Contexts

The fact that email headers, bodies and attachments may contain evidence of discussions and decisions makes messages very interesting to government officials, lawyers, and anyone else trying to study the past, ensure accountability, or even uncover misdeeds. Therefore, the legal context and policy environment in which email messages subsist should be of significant concern to anyone seeking to preserve them.The ways in which law and policy play out at the local level can lead to the destruction or preservation of messages that may have historical value, so archivists must be attuned to that environment and ideally, operate within it to ensure that email messages can be preserved.

If email often contains evidence and information that hold long-term value, why is it so difficult to preserve? At least part of the reason lies in the fact that the legal and regulatory regimes under which email messages are sent, received, stored and managed encourage risk management strategies that lead, at best, to passive neglect and, at worst, to active destruction.[4] In this respect, four areas of law and regulation have been particularly influential:

- **Public records and Freedom of Information (FOI) laws**: Public records laws establish or imply that email sent or received by public bodies is potentially a public record. It therefore must be managed in accordance with the principles of the prevailing law and best professional practice (Pennock, 2006, p.10; Baron, 2010). FOI laws, which affect many public institutions, might be seen as encouraging email preservation and accessibility, at least for the short term.
- **Financial oversight laws**: These outline records retention compliance periods and set strict rules regarding data use, compared with the FOI presumption toward retention. For example, US accounting law requires that accounting firms keep audit records, including correspondence, for seven years (United States Securities and Exchange Commission, 2003). Once defined periods have

[4] Maureen Pennock has provided a detailed overview of the UK legal environment (Pennock 2006, pp. 11–14).

been met, any further retention presents a discovery or privacy risk to the organization holding such records (Scholtes, 2006a).

- **Rules of Civil Discovery**: The US Federal Rules of Civil Procedure (FRCP) define the concept of Electronically Stored Information (ESI) and establish rules under which ESI must be provided during civil discovery process (US Supreme Court, 2010). In cases where a defendant does not produce ESI in compliance with FRCP requirements, or cannot show that the record-keeping system was maintained with integrity, the US courts can impose severe sanctions, leading to the emergence of an entire field of law and practice: e-discovery. The equivalent rules for the UK are far less prescriptive (Foggo *et al.*, 2007; Ministry of Justice, 2018). Overall, retention and discovery needs have encouraged an explosive growth in the market for email archiving and e-discovery software.
- **Data protection and privacy laws**: Data protection and privacy laws can also serve as an inducement for institutions to discard email, including those of potential long-term or historical value. The General Data Protection Regulation (GDPR), for example, was not written with email archives in mind, and its ultimate effects on records managers and archivists in the EU have yet to be fully determined, particularly when considering semi-structured data such as email accounts and messages. For many organizations, the arrival of GDPR raised the visibility of information governance and records retention as a strategic need. Overall, GDPR raises the potential for what the Archives and Records Association delicately called 'inadvertent outcomes' for the recordkeeping professions and function (Archives and Records Association, 2017). This is because GDPR includes provisions such as the requirement to consult data subjects and the right of erasure for those who request it. Formally the decision to preserve organizational emails is not affected, and people can still donate personal emails to a repository, like any other record. Nor does GDPR change basic practices, such as securing a deed of gift for emails that compose part of a personal archive; if anything, it reinforces their need. But will institutions be advised to consult third parties whose emails end up in organizational or personal archives?

What is the consequence of these four factors and in particular of the questions that GDPR raises? Government agencies and other bodies with public funding (such as many universities) have an interest in the long-term management of records, not only for their legal value, but also for their potential historical or cultural uses. Policies mandating the retention of records can be formalized into a records schedule, which may include review by an archivist and some attention to future research value. Private agencies (such as corporations or non-profit agencies that do not receive public assistance) also have an interest in long-term preservation for historical or administrative value. In the short term, archives should clearly articulate the public purpose of keeping email archives, advocating for their institutions to authorize a mandate to preserve records of historical value, if such an imperative is not already explicit (Archives and Records Association, 2017).

## 4.2. Email in Organizations

After gaining an understanding of the legal and policy factors discussed above, information professionals can begin to define the elements of a workable email management and preservation policy. This policy must also pay heed to the local institutional culture, which will shape the dissemination, management, and storage behaviours that email users exhibit.

Through policy setting, procedure development, and system implementation, an institution can either help or hinder the cause of preservation. In particular, perceived storage costs and legal risk can be barriers to effective email preservation, at least in some organizations.

Records managers, archivists, IT professionals and lawyers have in the past suggested that institutions establish policies and procedures so that end users of email systems identify and keep those records requiring long-term management, while trivial or 'non-record' items can be deleted. In practice, it is very difficult to set

up segregation procedures that people will consistently follow. In addition, any policy and procedure regime that allows people to delete individual emails permanently will simultaneously enable the potential destruction of records of long-term legal, administrative, cultural or historical value. At the same time, there is a pervasive perception that a 'keep it all' policy will induce headaches for current IT staff charged with storing the ever-growing bulk of messages, and leave the archivist a problem of even greater proportions: an unsorted mass of messages that—it is often assumed—will need to be manually disentangled and sorted.

This perception lies, in part, the rationale behind the US National Archives' Capstone option: that is, selection of 'archival' emails by capturing the entire accounts of people with a decision-making function within an organization. As the report of the Email Archives Task Force notes, even large well-established organizations struggle to implement formal records management programs for email. Most attempts to require users to capture and preserve email messages in EDRMS systems have not proven effective (Howard, 2011; Lappin, 2011). On the other hand, Case Study 7.1 shows how one organization is using semi-automated processes to capture, classify, and apply retention periods to email. And Case Study 7.2 demonstrates a method by which email can be bulk appraised at the end of a records system's lifecycle. Both of these options can be used in place of, or as a supplement to, Capstone methods.

## 4.3. Email in Personal Accounts

Personal information practices often prove more significant in ensuring email preservation than legal factors and organizational policies. As an essential, but often overlooked or even denigrated, tool in communication, email is easily lost, discarded, or deliberately destroyed. Yet people and organizations can take some specific actions to make the preservation of email more likely.

Of course, many people use and manage personal email accounts completely outside of an organizational structure or records management guidance. Archives will be interested in acquiring collections from people who lead interesting or influential lives, or whose experiences reflect social, economic, and political change. In these instances, email will be only one of many types of information to assess and select, as part of ongoing conversations. By forging relationships with such people, archivists can offer advice and assistance in preserving records that they might find valuable for their own purposes, pointing them to resources to encourage good email hygiene. As noted in the first edition of the report, many people treat their email as a de facto archive of sorts, using their accounts to send themselves documents or retrieve old photographs and reports.

Even if email records cannot be acquired immediately from the potential donor, it pays to build trust, since records that a potential donor preserves for his or her current use can be assessed by an archivist at a future time. When the time to donate arrives, archivists should discuss capture and appraisal methods. Tools such as ePADD (Email Process Appraise, Discover and Deliver), which came into maturity during the past several years, can be used to both capture and assess email for potential donation. Using software such as EPADD, users can connect directly to the source server and copy messages into the appraisal tool, or they can upload exported email messages. In other cases, an email account manager can export a PST file (from Outlook) or an MBOX/EML file (from other clients). Disk imaging and other forensic approaches may lead to the discovery of local copies of email.

The upshot: email preservation does not lend itself to a simple one-tool approach. Archivists will need to first identify the source email format, then choose from a suite of tools to move it into a processable format, such as MBOX. In parallel with this technical work, a deed of gift or other legal instrument should be used to provide the repository with clear title and rights to hold the email. And finally, clear access policies and procedures should be established upfront and made known both to donors and potential users.

## 4.4. Technical Factors Impeding Preservation

Given the legal, institutional and personal contexts within which email accounts are created and used, archivists and records managers need to understand a few technical factors that will affect their ability to capture, preserve, and provide access to email.

Most fundamentally, email records originate from what might best be termed a helper application. Email allows people to send any type of digital information, including information generated using other applications, from one email account to any other email account. In other words, email programs are simply communication utilities that support activities undertaken in the course of fulfilling daily work duties or in our personal lives. As a result, a single email account contains records of disparate context, structure and content, documenting activities both mundane and extraordinary.

Simply capturing and preserving the bits that comprise a message is straightforward enough, but further steps are required if the entirety of the message, including attachments or linked content, is to be accessible in the future. Since each email message includes a small amount of structured data (the header) along with a mass of unstructured data (the body and the attachments), preservation actions, if taken to an extreme, can entail a degree of complexity far beyond those necessary for preserving a homogenous set of files.

In order to understand the basic technical issues at play, we must understand how a wide range of hardware and software systems interact to send, receive, and store messages. In simple terms, email is a store and forward technology, centering on message transfer agents (MTAs) and user agents (UAs).

Nearly every modern email server operates in a way that complies with protocols and rules defined by working groups of the Internet Engineering Task Force (IETF). Messages can therefore be captured with relative ease by the receiving application or by a third-party service at point of transmission or receipt. However, once the message has been received by an MTA, the email server can do whatever it wants with it; it does not need to use a prescribed storage format. Further complicating matters, the email envelope, headers, bodies, and attachments may be received but not necessarily preserved in toto. (Section 3 of *The Future of Email Archives* discusses these matters in much more detail.)

From the end user's point of view, MTAs do not necessarily interact with UAs in a predictable fashion. At the moment a message is sent or received, the email servers or client enforce configuration directives and message handling rules which have been defined by the system administrator, the end user or both. Individual commands submitted by the end user (such as 'send', 'delete' and 'move' operations) are processed according to those rules and directives, but MTAs and UAs can interact with each other in ways that are difficult for the casual user to understand. Settings on the client and server may conflict, or the end user may not be aware of how his or her settings will interact with those recorded by the server administrator. As a result, email systems can facilitate what has been termed a 'corporate infestation', as messages replicate or disappear (Buckles, 2011).

A single message may be stored in many locations: on the server, on handheld devices, in local library files, on local file systems, on networked drives, and on backup devices. In spite of this replication, email is extremely susceptible to loss through deliberate action, user error, or malfeasance, since an action on any one device affects the master on the server, which then pushes the changes out to every client device. Looking at the entire ecosystem within an organization (as opposed to just a single account), any programmatic attempt to preserve email must begin not only with an understanding of the specific technologies used in an email network, but a detailed knowledge of how server administrators and end users have configured software and hardware. For example, some organizations may mirror, or journal, each message to an external store as it is sent or received.

In addition, any email preservation program needs to address other technical factors affecting the long-term usability of captured messages:

- **Context/Threads**: When users respond to a particular email message, they may not include all the relevant information from that message. Sender or recipient information may be a mere stub. In addition, different email servers thread related messages using variant protocols, some using nonstandard header syntax to link replies to a parent message. Reconstructing an individual message's context can require reverse engineering the email chain, or even the entire system. At minimum, it requires capturing current directory information so that those using the email in future years can establish the provenance of a message (Yeh and Harnly, 2006).
- **Attachments**: It is relatively easy to preserve the bytes that make up an attachment, either as they were encoded upon sending or in the original binary format. However, it can be difficult to locate message attachments, and email migration tools may not find them at the time of migration. In addition, attachments are potentially subject to format obsolescence. Email accounts do not segregate attachments by file type, and migration tools do not automatically filter attachments into different storage locations by file type. If an institution wants to preserve attachments, long-term preservation actions must be deliberately planned as part of a broader digital preservation strategy.
- **Embedded References**: Many email messages contain embedded links, referencing external files stored in another location. For example, emails may contain content found at a URL on a local or remote network location. In certain situations, it may be necessary or desirable to capture content at these locations, provided the content is judged to be a significant property of the message itself and is not available or documented via other methods. Obviously, such work would require significant forethought and technical sophistication, likely using a web crawler. When such records are stored in a cloud storage service, possibly behind an authentication barrier, challenges increase astronomically.
- **Communication Dispersion**: Given the now ubiquitous use of handheld devices, email is just one of many communication methods. While not yet a legacy application, attempts to capture it will, in most cases, not include the entirety of person-to-person messages that shed light on a person's or organization's activities. Therefore, other formats should be assessed alongside email.

## 4.5. Technical Factors that Facilitate Preservation

Email's core features include global addressing, interoperability, asynchronicity, redundancy, dispersion, backward compatibility, and extensibility (Task Force on Technical Approaches for Email Archives, 2018, pp. 24–25). Despite the technical challenges noted above, efforts to capture and preserve email are facilitated by the facts that the message exists in a standardized format at the point of transmission. This makes the main body of email messages (excluding attachments) into a near preservation-ready format (Dollar and Ashley, 2014).

The ways in which MTAs and UAs operate mean that the complete content of individual messages, including headers, bodies and attachments, can be captured by a third-party application at that point in time using email journaling applications, offering a real-time capture of all sent and received traffic. In short, because email was designed as a command and response system, messages are designed to allow for the easy flow of messages between different system actors, each unaware of the particulars of how the external system stores data.

Email capture is facilitated by another fact: essential header values for received messages are recorded in a structured, well-documented fashion. Each time the user hits the 'Send' key, his or her activity is documented as header metadata showing exactly what was written, to whom, and at what time. Specifically, all email messages must be structured with a mandated syntax for the message headers and body, in ASCII format. Although header metadata can be forged or modified after the fact (if the user has a deep familiarity with email format and some advanced computer skills), email is stored on central servers. Most end users would find such tampering difficult to execute and would likely leave a trail of their actions.

The increased use of server-based storage and the adoption of cloud-based email services such as Gmail and Hotmail also serve as trends that the archivist can use as leverage in the battle to capture email. There are several reasons for this. First, users can establish connections to these source accounts and export files from them in a way that they can be processed outside the email environment. This means, for instance, that a user of Microsoft Outlook can, using a few relatively simple instructions from an archivist or records manager, provide a version of the account, or selected portions of it, for preservation. Second, messages from the server are replicated outward, provided that IT managers have not configured the system with hard storage limits or rigid auto deletion policies. This leads to the widespread copying of messages onto local or client devices. If the archivist or current user cannot establish a connection to the source server, messages may be located in these other locations, stored in the format used by that application. Messages can also be captured from these locations, either by copying the files directly or by converting them from a disk image that may have been created as part of a larger digital accession.

Once it comes time to capture, sort, process or transform email, records managers and archivists can avail themselves of some recent advances in communication and preservation technologies. Each of these takes advantage of email's essential status as a standardized communication mechanism, and are treated in more detail in the following pages of this report:

- the emergence of email-specific application programming interfaces (APIs);
- the growth of forensic approaches such as disk imaging, toward capturing digital records;
- the emergence of natural language processing and machine learning tools for sorting and classifying messages;
- the continuing growth of emulation.

While the tools that implement these concepts may require advanced computer skills, they can be implemented as an extension to, rather than a replacement of, the basic capture and preservation methods that take advantage of email's standardized transportation and exchange formats.

## 4.6. Repository Perspectives

When considering these points, repository staff will need to take many factors into account. One helpful place to start is an email preservation readiness assessment.

This might be modelled on general digital preservation assessments, such as the Digital Preservation Capability Maturity Model (DPCCM) (Dollar and Ashley, 2014; Digital Preservation Coalition, 2017; Northeast Document Conservation Center, 2007), but it should include email-specific elements. Repositories should take, at minimum, the following factors into account when deciding how to assess, capture, process, store, preserve and provide access to email:

- Which general stage of preservation maturity does the repository consider itself to have met – nominal, minimal, intermediate, advanced, or optimal – in the terms of the DPCCM?
- Does the repository have sufficient legal authority or an institutional mandate to preserve email?
- Will the repository's general preservation policy and strategy need an adjustment to allow for the ingest and preservation of email?
- What source formats does the repository expect to receive? In what target formats does it expect to store preserved emails?
- Will staff need any specific training in email preservation, such as the tools mentioned in this report?
- If email containing private or sensitive information is being acquired, how will the repository identify such information? Will it be removed from the email collection, and if so, how?
- Will the repository's access policies and procedures need to be modified? Will specific policies need to be developed for collections that contain email?

Depending on the source of the email targeted for preservation, repository staff should consider a few additional factors. In the case of organizational records, repository staff should determine how email of longterm legal, administrative, historical, or other archival value will be identified:

- Can current records schedules and disposal information be applied to existing accounts?
- Should records managers and archivists consider implementing a Capstone-like policy?
- Which stakeholders should be consulted and involved in remaking policy (Watson, 2017)?

Overall, there are many challenges to address when moving organizational email preservation from a reactionary policy basis to one based on good information governance (Marshall, 2017). That said, Case Studies 7.1 and 7.2 show how two repositories are translating specific preservation needs into improved email governance.

In collecting repositories, different factors will come into play.

- Do deeds of gift or other acquisition policies need to be adjusted?
- Will the privacy of third parties need to be protected?

However challenging email preservation may seem, the issues noted above do not mean that email preservation is impossible, or even difficult. As the following sections demonstrate, the march toward maturity for email capture and archiving programmes provides information managers with powerful tools, provided we know to take advantage of them and can fit them into an effective policy and governance framework.

# 5. Standards and Best Practices

Developing an effective policy framework requires a solid understanding of email preservation standards, tools, and best practices. By fitting them into the structure of an overall programme and workflow, individuals and institutions can lay the foundation for a project or program to capture, store and manage email for long-term preservation.

Before beginning a preservation program, one must understand two basic technical details regarding the sending, receipt and storage of electronic mail messages:

- 1) email transmission and receipt is completely standardized around an open standard; and
- 2) processes surrounding storage and display are standardized only to a very limited extent. Each client and server implement storage and viewing in slightly different way.

The ways in which these two facts play out in a local context will shape the decisions that an information professional will make when implementing preservation tools and services. We can take some comfort in the fact that even though each system stores email a little bit differently, there are multiple pathways to convert email from one form or another. In principle, this means that after selecting a preservation target format, staff can design effective workflows to acquire, process, preserve and provide access to email.

## 5.1. IETF Standards

For the first 15 years of email's existence as a communication technology, the format of an email message was not standardized. Institutions and projects used a variety of competing methods to send, receive, and store messages (Partridge, 2008).

Since 1981, the Network Working Group of the Internet Engineering Task Force (IETF) has defined the methods that may be used to send and receive messages, using a series of 'request for comment' (RFC) documents. The *Future of Email Archives* (Task Force on Technical Approaches for Email Archives, 2018b) offers an excellent overview of these standards.[5] Information professionals and organizations seeking to preserve email do not need to fully comprehend these technical specifications, but by gleaning a few simple facts from them, they can make better decisions about the appropriate tools and services to best facilitate capture, processing and storage. The following points are particularly relevant to information professionals, since they shed light on potential capture and preservation pathways:

- **Message Transport**: The method by which mail is sent and received is defined by STD 10, the Simple Mail Transfer Protocol (SMTP) and, somewhat confusingly, by RFC 5321, which has never achieved the status of full standard but is widely implemented (Postel *et al*., 1982; Klensin, 2008). Roughly speaking, they define the 'envelope' within which each message is sent. The envelope is typically discarded upon successful delivery, so it usually plays no role in email preservation, unless a repository is able to use email journaling software to capture a message at the point it is sent or received.
- **Message Structure**: RFCs 5322 and related standards describe the Internet Message Format (IMF) and Multipurpose Internet Extensions (MIME), defining requirements for the format of the contents of that envelope (Resnick 2008; Brodkin 2011). While receipt systems typically discard the envelope, they retain the full header and body, perhaps in a format based on the IMF requirements, or perhaps not. The precise storage format is left to local implementation, but messages must be transmitted between servers using SMTP and IMF. At the time a file is stored locally, the system may add additional metadata, such as spam scores, the results of authentication processes, and user

[5] Tobias provides a more concise, non-technical summary (2017), and Task Force on Technical Approaches for Email Archives (2018a) provides a more detailed view.

classification information, such as tags and folders. This means that the stored header and message, including MIME headers, will include information that can help users judge the authenticity of

messages. Ideally, any information professional will preserve all of the metadata from the header during the email capture, processing, and preservation storage. Working with programs that can migrate email using native formats helps ensure authenticity. Processes that create a derivative file, such as a PDF, may lose some or all of the metadata that can be used to authenticate a message or set of messages. It may also result in the loss of functionality for attached files or other content, if they are not preserved in their original format.

- **Message Retrieval**: When a message is retrieved from an inbox or other storage location, the operation is often executed using the Internet Message Access Protocol (IMAP) or, less frequently, the Post Office Protocol (POP3). These standards specify how user agents may connect with email servers to allow an individual to view, create, transfer, manage and delete messages (Crispin, 2003; Myers, 1996). Most email servers support one or both of these protocols, since doing so allows users to connect to the server using the client application of their choice. The digital preservation community can leverage this factor, and in fact several of the tools for email archiving and preservation allow users to connect the tool to accounts using the IMAP protocol.

## 5.2. Implications of the Messaging Model

Figure 1 provides an overview of the email messaging model as specified by these standards, developed by Joel Simpson for *The Future of Email Archives* report. Information professionals should have a working knowledge of the transport, structure, and retrieval process that is implicit in the email messaging model. Some (but not necessary all) of its features will be supported by particular client/server combinations, and the model can be extended with proprietary additions. Thus, information professionals must understand which client/server combinations are supported, if they hope to intervene. Depending on the configuration, particular steps will need to be executed if the repository hopes to capture and store messages. The specific decisions taken (to use or not use certain tools, in a certain order) will impact the specific workflows and architectures that best serve preservation interests.

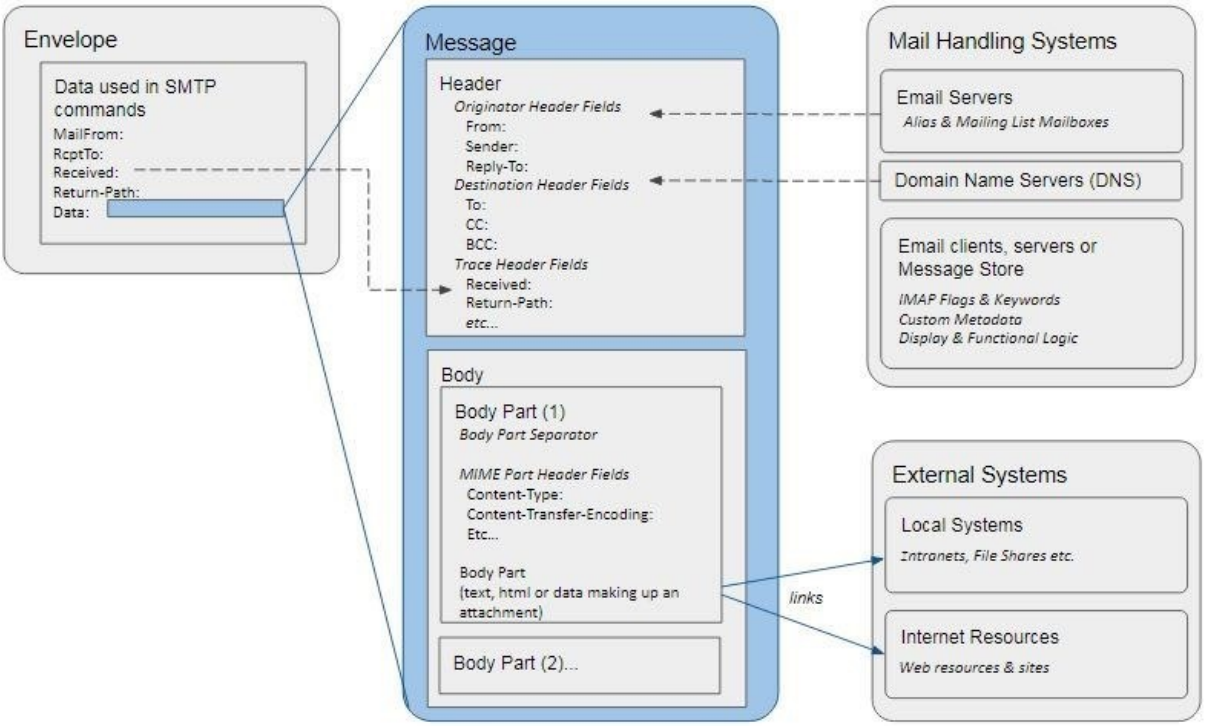


*Figure 1: Email Message Model (Task Force on Technical Approaches for Email Archives 2018b, p.27)*
*Graphic developed by Joel Simpson and copyright ©2018 Council on Library and Information Resources. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International Licence.*

Today, nearly every email server and client supports this messaging model. This includes proprietary software such as Microsoft Exchange, Lotus Domino, and Novell First Class, as well as open-source servers such as Postfix, Sendmail and Gmail. Although we do not need to delve into additional detail regarding these standards, several related facts are worth noting:

- The Internet Message Format (IMF) serves as the basis for two of the most common storage and exchange formats: MBOX and EML (Hall, 2005; Wikipedia, n.d. 'Email'). Since neither of these transport standards mandates a storage format, each server or client is free to store messages using its own internal logic. Yet many systems either use one of these formats as its storage method, or make it relatively easy to convert messages into them. MBOX (sometimes known as Berkeley format) is a set of four slightly different storage formats, developed originally for Unix systems. Generally, a single file with the extension .mbox or .mbx contains the contents of an entire folder, including any attachments embedded in MIME. Files can grow to astronomical sizes, and even slight file corruption may affect the ability of certain email client software to access individual messages or even the entire folder. EML files, on the other hand, typically include each message as a single file, and attachments may either be included as MIME or written off as a separate file, referenced from a marker in the EML file. In spite of these issues, MBOX and EML have achieved a certain status as de facto standards. In the case of some proprietary clients, messages cannot be exported from their native system directly into MBOX or EML. Instead, these clients may export the message to a proprietary, though perhaps open, format. The most common of these formats are .pst (Outlook), and .nsf (Lotus). Tools such as those discussed later in this report can convert these files to MBOX or EML. Similarly, an institution might come across email that originated in an obsolete system. In this case, tools such as Aid4Mail, Emailchemy or Xena can convert many file types to MBOX or EML formats. Institutions may need to exercise creativity in migrating files if standard tools do not support them, or they may need to use relatively expensive forensics software to access the files. In

general, if an institution can convert email into one of the MBOX or EML formats, it has taken a very big step on the road toward preserving email, perhaps even a sufficient one if the goal is simply bitlevel preservation.

- Particular server–client combinations also support proprietary protocols, such as the Message Application Programming Interface or MAPI (Caputo and Narva, 2016); similar protocols exist for IBM/Lotus Notes/Domino. In addition, as email services move to the cloud, service providers are beginning to provide stateless Application Programming Interfaces (APIs) as an alternate means to access and process email via HTTPS ('JSON Meta Application Protocol Specification (JMAP)', n.d.). These developments should be monitored closely by the preservation community, since they may provide new means to capture email data.
- Some servers use alternate methods to supplement, extend or replace functionality that is specified in the IETF standards. Typically, these features are added by defining extended headers, which are then manipulated using proprietary message access protocols. For example, most IMAP-compatible servers support the message-id and in-reply-to fields in order to track and reconstruct emailthreads, but Microsoft Exchange servers supplement them by including a thread-id header and many other headers intended to facilitate message reuse and to allow for server-specific features such as spam detection.
- Authentication technologies have been added to the IETF standards over the years; these include the Sender Policy Framework (SPF), Domain Keys Identified Mail (DKIM), and DMARC. Where digital signatures exist in an email header, they should be preserved, because they will allow users to make stronger claims of authenticity regarding a particular message, even if such claims cannot be made with absolute certainty. (For example, future users may find that a referenced certificate has been removed from the remote server that issued it.) Nevertheless, documenting the chain of custody and preservation can lead to increased trust in a message's authenticity.

- Removing messages from their native systems may negatively affect people's ability to search, discover, retrieve or render messages, and may lead to the removal of some header or metadata information. Additional testing is necessary.

## 5.3. Evolving Practices

As noted above email servers and clients frequently use the standards-based MBOX and EML formats, as well as the proprietary but documented PST format, as storage and exchange mechanisms. Email preservation tools also use them for import and export, less commonly for internal storage. But many tools make use of other propriety storage options. Taken as a whole, these facts open up three potential preservation approaches for email: bit preservation, migration, and emulation.

- **Bit Preservation**: Many archival repositories will be well positioned to undertake a basic email preservation project by capturing and storing messages in whatever format the external system provides them. For example, a repository may simply keep a set of emails in proprietary but open and documented format such as PST. While these strategies would leave many unresolved questions (such as how to preserve and render attachments over the long term), they would provide a baseline of preservation, allowing for the development of more refined practices in the future. Optionally, attachments could be converted from MIME to their original binary formats, then stored as part of an archival packet, alongside the original data object in the source format, and optimally with a pointer back to the original message.
- **Migration**: With just a pinch of additional effort, information professionals can convert captured email messages to open formats, such as EML or MBOX, then preserve them within a digital repository. Again, this approach will not solve all preservation problems. However, it will alleviate dependence on a particular platform and allow for the movement of the messages between particular tools, such as ePADD, that can filter, sort and redact messages. In theory, migrating

messages to an XML format could help facilitate the long-term preservation of messages. The Email Account Schema, an XML format developed by David Minor and Steve Burbeck, provides an excellent initial implementation of such a storage format, since it puts email into a self-describing format, using key–value pairs. Their work has been brought forward into the TOMES — Transforming Email through Embedded Semantics — and DArcMail projects (Smithsonian Institution Archives, 2008; North Carolina Department of Natural and Cultural Resources, n.d.; Watson, 2017). A few tools can now capture, filter, query, display, and render email messages that are stored in the XML format (Smithsonian Institution Archives, 2017; Gibson, 2018). Standardization around the Email Account Schema would be greatly facilitated by the development of additional applications that allow users to search, discover, view and visualize messages that are stored in the XML format. In the meantime, institutions that decide to keep email in an XML format should also keep a copy of messages in one of the IETF formats, pending the development of additional rendering options for the XML-ized version of the emails.

- **Emulation**: Sometimes, an information professional may decide that an email collection would be best rendered and used in its original environment. In these cases, the institution will sometimes confront an initial challenge: determining exactly what composed that original environment. Assuming that this can be worked out and the appropriate software located, the collection may be loaded into an emulator, where it can be browsed, accessed and viewed using the same tools that were available to its original creator (Carroll *et al.*, 2011). Although emulation is a complex process, there is some hope that it will become increasingly feasible over time to provide access to emulated versions of email kept in its original environment, at least where a user accessed the message with a specific client application. (Emulating email environments for those who used multiple devices or browsers apps would be less likely.) At this time, projects such as the Software Preservation Network and the Scaling Emulation and Software Preservation Infrastructure (EaaSI) program should be monitored carefully and considered for potential use, if a repository wishes to preserve the original look and feel of an email reading and browsing experience.

## 6. Technology and Workflows

The standardized way in which email is transmitted and stored, as well as the potential migration and emulation practices noted above, give repository staff many options to consider when pursuing a programme of capture, processing, and preservation. General guidance from the digital preservation community can be used to develop a baseline understanding before undertaking email preservation (*Digital Preservation Handbook*, 2015; Marks, 2015; Owens, 2018; Purcell, 2016). In general, people wishing to preserve email should become skilled in using multiple email clients, such as Thunderbird, Outlook or Apple Mail. Experience with web interfaces (such as Outlook 365's archive or Gmail's Takeout feature) is also helpful. By practising with such tools, archivists build a skill set for working with more obscure or defunct email clients.[6] Critically, staff should be skilled in using software to import and export messages, as well as in connecting to accounts using the IMAP protocols.

Once such skills have been developed, several types of tools can be used to support email preservation work, in three functional areas: 1) Capture; 2) Appraisal and Processing; and 3) Search, Discovery, Access and Rendering. The remainder of this section covers those topics, before briefly charting some sample workflows. The case studies in Section 7 should prove helpful for repositories looking to design their own email preservation programmes.

[6] Wikipedia's list of such clients can be very helpful in getting one's bearing or defining migration pathways (Wikipedia, 2018a).

## 6.1. Capture

The first and, in some ways, most pressing problem for those wanting to preserve email is how to acquire it in the first place. As the foregoing discussion notes, each email represents a transaction. People (as well as the servers and clients that they use) treat it as a stream of information that is, potentially, written to one or more locations. Message transactions are most complete at the moment of sending/receiving, and some information (for example, the envelope) is lost when the message is written to storage. At the same time, other information (for example spam scores, authentication results) is added. Some users delete messages as soon as they are finished with them, a policy encouraged by zealous retention/deletion advice or allowed by lax information governance regimes. Other users (academics are notorious for this) allow email to indefinitely accumulate. They may dutifully put it in folders or, like the author of this report, let it agglomerate in an inbox, relying on an application's native search functionality to provide an occasional window into their past. Many people treat it as a de facto, albeit informal. archive, retrieving messages and attachments from their history when needed.

For practicality, the capture processes that a repository institutes must sync not only with institutional policies, but with the user behaviours that those policies encourage or tolerate. It must also be trusted by the people whose email will be captured. More often than not, capture strategies will be aimed at accounts, or portions of them, rather individual messages.

Given this general (albeit vague) advice, managers should: 1) pick one or more of the three general capture strategies listed below; 2) seek to embed it within an information governance framework; and 3) instrumentalize that framework into specific implementation policies and tool choices.

- **Migrate email from one storage location and format to others**: this step might be an easy win, at least if the account owner still has access to the account and can follow a few simple instructions. Most modern email servers and clients provide the ability to export messages into a format that other systems can read, such as MBOX or PST. Using features such as the 'Export Mailbox' feature in Apple's Mail Client or the instructions provided by Microsoft, people can download a packet containing an entire inbox or portion thereof (Microsoft Corporation, n.d.). But what should an archivist do if the account can't be directly accessed? Email migration software, such as the commercial software Aid4Mail and Emailchemy, as well as open-source, multi-purpose tools DarcMail and ePADD, will read email in one or more defined formats and save it in one or more defined target formats (Smithsonian Institution Archives, 2017; Schneider, Chan, and Edwards, 2017; Stanford University, n.d.). If the archivist possesses account credentials, they can attach such software directly to existing accounts and harvest messages from a server. If local copies of messages are found during the appraisal of some storage media (such as a hard drive image), they can be migrated from one format to another. The most sophisticated commercial software, such as Aid4Mail, converts email from many obsolete formats, connects directly to IMAP-compliant servers and includes both filtering and scripting functions to allow customized output. The output of these capture processes can be written to a secure location, allowing for appraisal and process steps noted in the following sections of this report.
- **Capture email at time of transmission or receipt**: this is an alternative approach, much more costly and likely to be more difficult to implement, for both administrative and technical reasons. 'Email archiving' tools capture the email text stream at point of transmission, saving it to an external store, typically a database of some sort, but not necessarily in MBOX, PST, or other transposable format. More sophisticated packages include the ability to filter messages on capture, to browse and search the store, and to apply retention periods and audit rules to messages. Examples of such software include Symantec Enterprise Vault, Smarsh, and Mail Archiva. The latter comes in both open-source and enterprise flavours. Email archiving tools such as these almost certainly require professional systems support and maintenance, and they are typically implemented in commercial businesses and governments to ensure compliance with records management needs. To the author's

knowledge, no published literature discusses the ability of vendor-provided email compliance systems to preserve email permanently, although it seems likely that systems which store messages in an open format would be more likely to allow for migration; such details would need to be reviewed carefully as part of an request for proposals or bid process.

- **Tools to manage email within an EDRMS or Content Management System**: Electronic Document and Records Management Systems, which typically provide an organization with a means to comply with records management requirements, may include a method to declare, register or classify email as records. Institutions pursuing the use of EDRMS software must plan to spend a considerable amount of resources acquiring, configuring and supporting the application, as well as defining and enforcing policies and procedures that facilitate effective system operation. Such systems can be useful in highly centralized or litigation-sensitive industries and agencies, but their efficacy in capturing archival email has been mixed at best, and has in fact been largely abandoned in favour of the Capstone method – capturing whole accounts that meet some functional appraisal criteria, perhaps at set intervals or at the end of an account's life cycle.

In short, those seeking to capture email must be aware of, and prepared to use, many tools, not wedded to a single solution. Whichever methods are selected, the goal of this capture phase is simple: to remove email from an environment where it is managed by the end user or system administrators. Those groups may have less interest in its long-term appraisal, than do records managers or archivists. Once the email is captured, it can be converted to a standardized format amenable to several workflows and to the types of iterative processes to which we now turn.

## 6.2. Appraisal and Processing

Since the first edition of this report was published in 2011, the community has developed both tools and work processes that can be used to identify, filter, sort, refine, tag, arrange, describe, and package email. Each of these steps typically takes place before an email collection is ingested and stored in an archival repository. A single approach won't meet all needs, and repositories should consider mixing open-source and commercial tools into a workflow pipeline.

On the open source side, several tools are particularly noteworthy and useful. ePADD, developed by Stanford University, and TOMES, developed by the North Carolina Department of Natural and Cultural Resources, include features that support natural language processing, authority reconciliation, tagging, and sensitive content review. They facilitate message removal or redaction and include basic export and packaging functions. DArcMail (developed by the Smithsonian Institution Archives and partially integrated into TOMES) provides a method to transform and package email in the XML-based EAXS format. And most of these tools provide ways to separate attachments and other binary content from the messages from MBOX or PST files, transforming them from MIME content into their native binary formats. This makes them amenable to file analysis and the development of preservation strategies as part of a repository's overall preservation planning.[7]

Figure 2 illustrates the main interface for ePADD. The archivist or account owner can use one of several filtering mechanisms to winnow a large corpus of messages down to a reviewable bite, then mark single messages or entire groups for inclusion in an archival packet, to be exported at a subsequent phase of an appraisal or processing workflow.

[7] For a more comprehensive set of tools see the *Task Force on Technical Approaches for Email Archives* website: http://www.emailarchivestaskforce.org/documents/email-tools/

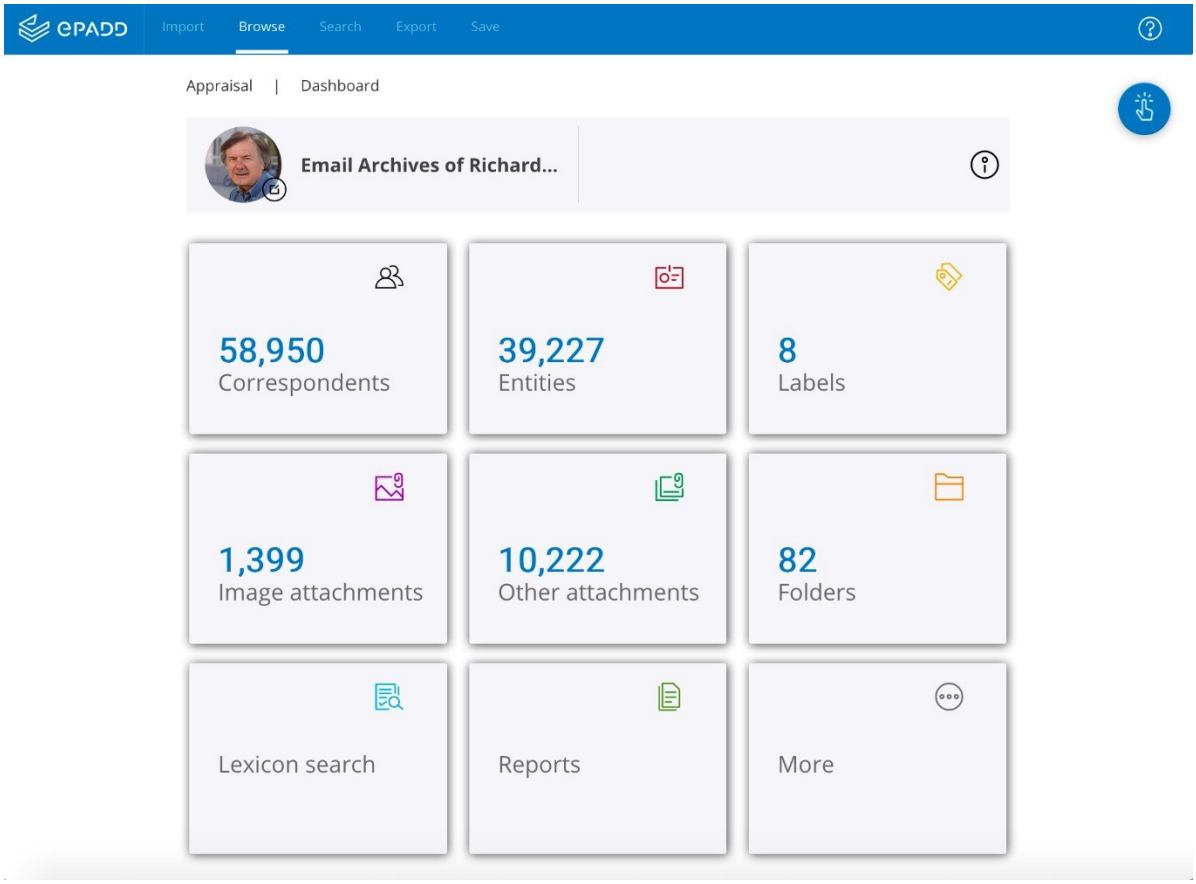


*Figure 2: ePADD Interface. Courtesy Stanford University*

The TOMES software was produced as part of a larger grant project and is aimed at processing large email accounts at scale. It includes software to convert emails from PST to EML format, and then to package content into EAXS, with messages tagged with entries from a named-entity dictionary. While the project developed a named-entity lexicon suitable to state government, TOMES also provides a method to define other lexicons.

The commercial sector provides tools that will also be of great value to archivists, and even when the heart of an appraisal workflow uses open-source tools, such commercial tools typically play a vital supporting role, most often in the area of migration and transformation. Aid4Mail, Emailchemy, and Read PST, for instance, are often integrated into archival appraisal and processing workflows.

In addition to these migration tools, some institutions have begun experimenting with commercial software that can be used to classify or sort email and other text-based content.[8] Most of this software uses machine learning algorithms or other advanced computational methods, which tend to operate a bit like a black box. As will be discussed in Case Study 7.3, the University of Illinois at Urbana-Champaign is experimenting with predictive coding software as a way to more easily segregate record from non-record materials.[9] The Virginia

[8] The need for these machine-assisted approaches was dramatically demonstrated during a US Supreme Court Confirmation Process, when the National Archives Review Process for email records broke into a minor scandal (Kim, 2018).

[9] Additional Information about this project: https://www.uillinois.edu/cio/services/rims/about_rims/projects/processing_capstone_email_using_predictive_coding/

State Archives, likewise, undertook a human v. machine contest, in which continuous active review software proved surprisingly effective (Cormack and Grossman, 2017). This approach, that of machine-assisted classification requiring human feedback and decision making, seems particularly encouraging and is the focus of continuing research work, building on the TOMES and Bitcurator projects.

Whether open-source or commercial tools are employed for appraisal, these filtering tools will typically be used to remove messages that record creators or archivists do not want included in either the archival or dissemination packet. Given the general sensitivity around email content, it seems likely, if not inevitable, that most email datasets will go through several rounds of review as software iterates through previously refined datasets.

Once such winnowing is complete, the same tools listed above can be used to package the files for ingest to a repository. The precise packet structure that each repository uses should be determined by local policies and the requirements of the receiving system. It will be particularly important to ensure that attachments and other binary content are extracted and saved separately from their parent messages, in a way that ensures the parent–child connection remains unbroken over time.

## 6.3. Tracking Processing and Preservation Actions

While an archive's staff may have relatively little control over external factors impacting message authenticity, they can take specific actions to demonstrate the archive's trustworthiness. For a typical digital repository, actions such as the following show considerable good faith in documenting a collection's provenance, ensuring its chain of custody, and tracking its processing history:

- registering the transfer of ownership or custody of the material;
- ensuring contextual information is retained; for instance, repositories should preserve attributes such as the folder structure of an email account, the relationship between emails and their associated attachments, the relationship between the email account and any other digital archive material being transferred, and additional metadata that might exist about the material;
- maintaining a full audit trail of any actions taken on the material, and the person or system responsible for carrying these out;
- running fixity checks on the material when it is copied or moved from one storage location or medium to another;
- recording repository actions as part of the preservation metadata that accompanies the email throughout its life.

As this list implies, the principle of provenance can be applied to email without significant deviation from that applied to other digital formats. That said, most archivists lack access to a full-blown, integrated toolset to support active, ongoing documentation about the actions they perform on email corpora. Institutions

without automated systems can manually record information, then store it electronically in their collections management database.

For some or all these actions, a large institution may have access to a mature preservation repository infrastructure that can record the results of automated processes. Ideally, these records will become part of the digital provenance and help system administrators track digital objects for long-term preservation and information management.

## 6.4. Preserving Attachments and Linked Content

Attachments pose particular challenges for the archivist processing an email collection, beginning with the seemingly simple task of ensuring that an attachment remains linked to its original message and that it can be

rendered for viewing. Less immediately obvious, but even more problematic, is content that is linked to a message through a URL. The URL might be visible to the end user, or it may be embedded in a way that makes it appear to be an integral part of the message, even though it is actually stored on a server distant from the main body of the email message.

Archivists must address the following areas when working with attachments and linked content:

**Potential for Loss or Corruption During Email Conversion**: throughout the email workflow, conversion tools may fail to properly handle attachments, resulting in complete loss or corruption from an incorrect conversion. When capturing emails and preparing them for processing, it is often necessary to convert from proprietary email client formats to a standard format such as EML or MBOX. Similarly, certain client or server applications may export emails in a way that does not fully preserve attachments. This issue can be manually addressed by monitoring the results of the conversion through the application of quality control procedures. The time- and resource-intensive nature of such work suggests that semi-automated testing regimes would better facilitate reproducible workflows within and across repositories.

**Processing Issues**: when working with email collections, archivists and curators should be cognizant of several risks. By carefully attending to the following factors, they can ensure that the preserved files are more authentic and useful to future researchers:

- **Potential deletion**: it is easy to accidentally delete an entire email when only the attachment is to be preserved. When an attachment is embedded, it must be extracted from the email in order to be preserved separately from the email message. This suggests that the processing history of attachments should be tracked in addition to the history of the message itself. Most critically, systems should record the removal or disassociation of attachments from the message. Automated removal of attachments especially comes into play when replying to or forwarding a message, so archivists should also be careful when deduplicating collections or threads.
- **Format issues**: attachments may be of many different formats, each potentially requiring a different tool for identification and evaluation. However, technical and structural review of attachments may be accomplished with commercial software such as Quick View Plus or FTK (which contains Quick View Plus).
- **File size**: some attachments, such as video, may be very large, placing a strain on storage systems and, where files are transferred across networks, on the network. Archivists should also be aware that some systems store large files externally at the time of mailing or receipt. Gmail, for example, can be configured to automatically store large objects in Google Drive and create a link in the message, while Apple's Mail Drop provides similar functionality (Gmail Help, 2018; Apple, 2018). Both systems introduce access and retention restrictions that might render attachments inaccessible in the future. Preservation is a constant risk with email systems, which routinely add new features to enhance the user's system experience.
- **Viruses and malicious content**: attachments may contain viruses or other malicious content. These must be detected and procedures put in place for determining how to handle such files.

**Repository/Preservation Issues**: Once an email account or set of messages has been processed, an archival repository must confront storage and long-term preservation issues, focusing on two themes:

- **Attachment storage and handling**: when an attachment is embedded within the email message in MIME format, it is relatively easy to guarantee that the attachment will remain associated with that email. However, attachments are easiest to monitor and store in their native binary format, not as MIME content embedded in a message. To ensure that the stored file can be located, a pointer should be placed in the original message. Many email client applications do this automatically, if somewhat invisibly. In the context of a digital repository, which must preserve content for long periods of time, there are many unresolved questions about how to maintain fidelity between the

message and attachment. But the upshot is that repository software should have a method for the persistent linking of attachments and email messages. At the simplest level, the relationship between a uniquely identified message and attachment could be documented in a spreadsheet or METS file. A more complex solution might entail modifying the body of the message to insert a pointer to the new file location. Any solution will require careful planning, as the storage location may be moved at any point in the processing workflow or during subsequent maintenance of the repository.

- **Migration**: files attached to emails take on entirely new format preservation issues over time when compared with the body of an email message, which is typically just ASCII or UNICODE text. This suggests that preservation policies based on email formats (for example, deciding on a target storage format for the message itself) will not adequately address the preservation of attachments over time. If attachments are placed in a repository, they can presumably be migrated to formats that constitute a target preservation format. But this raises a set of additional questions: will the original attachment be retained? How will the target file be associated with the message? Will formats be monitored over time, and if so, how will additional migrations be tracked?
- **Linked Content**: preserving linked content in email collections is a parallel challenge to preserving native content within the message or attachment. Many organizations now require or request that staff save files in file management systems (e.g., SharePoint or Box) and only provide a link to the intended content via email. In this case, email acts as the connection to the data but does not contain the data itself. While this is inherently a business decision – helpful for saving space and avoiding data collisions when many people are to approve or review the same document – it does have significant implications for archiving. In terms of benefits, the risk of exposure is lowered if the link leads to a Box folder requiring authentication. If the email message is inadvertently forwarded or the email account is hacked, the unintended recipient can't access the file. Yet security is a double-edged sword. Without the record creator's credentials or access to the folder, the archivist or user can't access the file either, perhaps long after any need for confidentiality has passed. For these reasons, the archivist may work with the donor to acquire a separate copy of the files, obviating any need to rely on an external system.

In other institutions, the decision to preserve linked content might be more flexible, informed by available resources and collection development policy. Since the linked content is stored outside the message, archivists may decide not to preserve it. Such a strategy has, of course, precedent in the analog world; archivists would not necessarily track down a publication or other document that is mentioned in a folder of correspondence but would simply assume that interested users could try to track down such documents as part of the research process. In some cases, however, archivists may decide that it is important to save such content, based on collection analysis or some other factor.

The presence of such linked content is increasingly problematic when integrating email with file management and messaging systems such as Sharepoint,[10] Slack,[11] Box,[12] Dropbox,[13] and Google Drive.[14]

The linked content may, in fact, be integral to, if not more important than, the message itself. As anyone who has tried to access a photograph or document while off the network knows, such content is simply inaccessible and will not render in the message if it cannot be located. With networked resources, these issues become more prominent over time: locations or content can change, or the content might simply never have been preserved. Referenced network resources may not be accessible if they are located behind a

[10] Sharepoint: https://products.office.com/en-us/sharepoint/collaboration
[11] Slack: https://slack.com/
[12] Box: https://www.box.com/en-gb/home
[13] Dropbox: https://www.dropbox.com/
[14] Google Drive: https://www.google.com/drive/

firewall, on an internal network, or in another location that can be accessed only by authenticated users. Some links have been run through a URL-shortening service/link management platform such as bit.ly.

Because these services produce a URL that may have no association with the actual location, even the most diligent researcher would be unable to definitively identify the original resource from the URL. Some, like bit.ly, promise to maintain the link permanently, but this will always be contingent on the continued existence of the service.

Email archiving applications have not, as yet, provided functionality to harvest resources at a URL that might be found in a message, much less to preserve embedded images, documents, or other content. A solution to mitigate the second situation would be for the application to recognize known services and follow the link to record the actual location of the resource. To do so, the application would have to be kept up-to-date with major services and would provide only an original location for the resource. Another, and more comprehensive, solution would be to link the email-archiving application to software that could retrieve the resource or store it at an allied location, such as a web archive collection or as a linked resource (similar to an attachment). For the sake of practicality, this would have to happen in a timely manner – such as in a message journaling workflow – to capture the linked content before changes occur.

The key issue is whether tools used in email archiving should extract web links, attempt to crawl those pages, create a unique identifier such as a DOI or some other persistent ID to the archived web page/site, then replace the link in the archived email with the new permalink. This question highlights the challenges of both context and scale for email archives.

Adding the linked content to the email collection would ensure that it continued to exist, though at the price of increased storage cost and complexity. Because it is often impossible to know the full extent of a web resource, web archiving scope could become an issue as well. Finally, it is hard to imagine what solution can be devised for situations in which the location is inaccessible because the resource has been moved or deleted, or exists within a private network, other than allowing a harvester to authenticate to that location with the provided credentials.

In short, accounting for attachments and linked resources in email collections is a difficult task, but one for which institutions and developers will have to account. The set of issues identified above suggests many potential avenues for development

## 6.5. Preservation and Storage Options

Cloud-based preservation services are beginning to offer email-specific preservation services, and developer documentation can be helpful when designing a specific deposit structure, regardless of the specific tool being used (Archivematica Wiki, 2017). Those seeking to work with a specific cloud-based system or other preservation services vendor should closely coordinate ingest attempts with support staff from the provider.

For those institutions using a home-grown or file-based approach to email archives, I suggest a variation on the simple packet structure shown in Figure 3. The particular elements of this, such as filenames, can be varied to suit local policies. Consistency is the key, and advice such as that provided by Tim Gollins is still valuable: we should aim to be ‘using only the minimum necessary intervention to secure our digital heritage for the next generation’ (Gollins, 2009, p. 75).

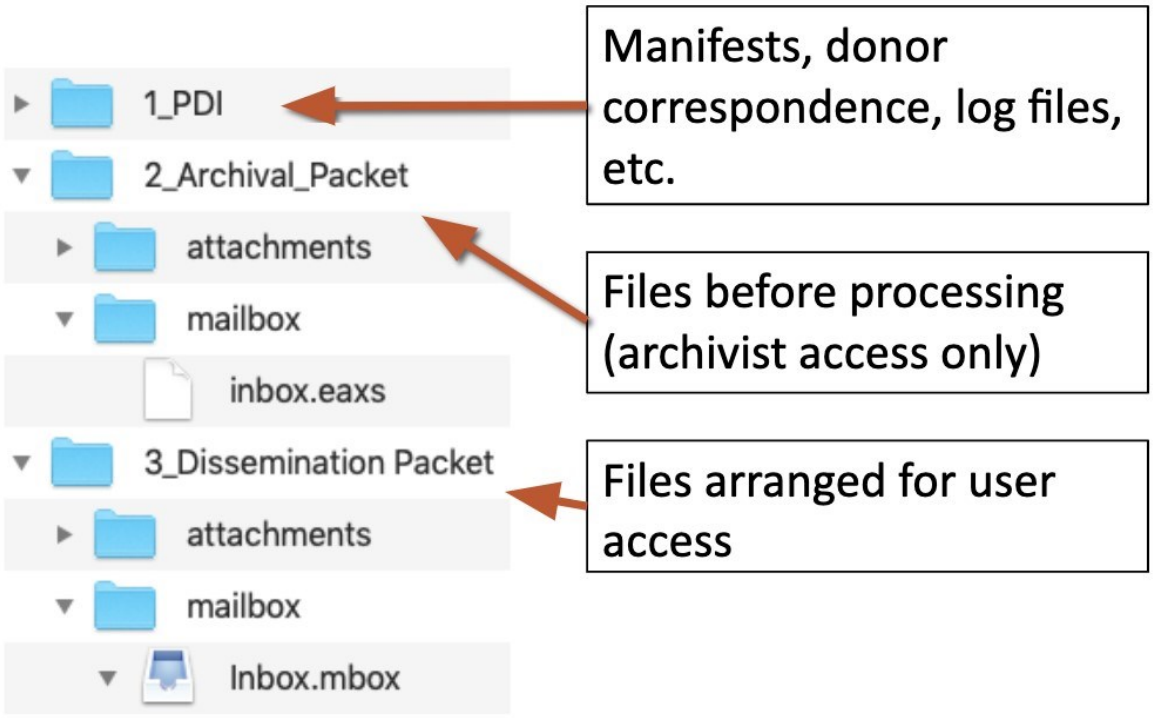


*Figure 3: Parsimonious Preservation Approach to Email Storage*

.

## 6.6. Search, Discovery, Access, and Rendering

Email that is being processed should, of course, be treated with an end goal firmly in mind: that of user access. For these reasons, a repository's search, discovery and access tools should be kept firmly in mind when defining the package structure and, in particular, the shape of the dissemination packet. Yet the future is far from certain, and search systems or discovery tools may change, possibly quite rapidly. For these reasons, materials should be packaged in a way that ensures their future accessibility. Archivists might consider, for instance, making the email collection available as a form of data which can be downloaded in a system-neutral format such as MBOX. Such a step would serve both present users and future archivists well.

Whatever means are used to disseminate and render messages and attachments, repositories should establish a baseline of good practice in the ways they describe the email packets. Item-level approaches hold some intuitive appeal and have proven influential in some contexts, although they can impose heavy costs, unless descriptive metadata can be extracted in an automated way (Bailey, 2013). In this author's opinion, they should be resisted, at least at first, when repositories are thinking about how to provide access to email. In most instances, a good series-level or collection-level description of the entire email corpus, including appropriate authority control, a scope contents and extent statement, as well as date range, processing notes, and right statements, will help the end user best assess whether an email collection is useful. Using guidance from archival descriptive standards such as ISAD(G), repository staff can construct simple and effective metadata records within their existing catalogue or archival management system. Since the messages themselves are already in text format, indexing tools or even email applications will provide more effective search, browse, and retrieval mechanisms than a finding aid or indexing in some non-email specific environment.

This is not to say that item-level metadata has no value. As noted above, tools such as ePADD and TOMES can be used to tag individual messages, in an automated fashion. Since they also allow for frequency reporting

and other quantitative measures, their output can help archivists identify particular important subjects, places, people and organizations. By setting a frequency threshold, or simply looking at the output of these tools, archivists and collections managers can elevate such terms and apply them as access points to the aggregate descriptive record.

Over the longer term, item-level metadata can be leveraged by email-specific access systems. At the simplest level, email files themselves contain a wealth of unstructured or semi-structured metadata. This metadata is used by email clients, such as Microsoft Outlook and Apple's Mail app, to provide sophisticated querying search and rendering features. By providing an MBOX file to users – that is, treating the email collection as data – archivists will let users leverage the power of the tools they are most familiar with. Users simply need to import the file to begin using it. This may seem like a low-effort strategy, but it seems unlikely the cultural heritage community will ever be able to create email access software quite so powerful as the tools developed by Mozilla, Microsoft, Apple, and Google.

That said, there will be many cases where it is neither feasible nor desirable to make an MBOX file directly accessible to users. In these cases, archives should use tools such as ePADD's discovery and delivery modules, which take full advantage of the item-level metadata produced by such tools. Longer term, the development of such solutions holds considerable promise. ePADD's discovery module, for instance, provides an emailspecific environment to search, browse, and use email messages, situating each message within the provenance and context of the corpus to which it belongs. Over the very long term, one can envision institution-wide or cross-institution databases of email correspondence, similar to the types of resources offered for analogue archives.

Up to this point, this section of the report has described the individual elements of an email processing workflow. Now let's put the pieces together. Figure 4 provides a simplified email processing workflow across the entire lifecycle of email accounts, from creation through archival acquisition to discovery and access.

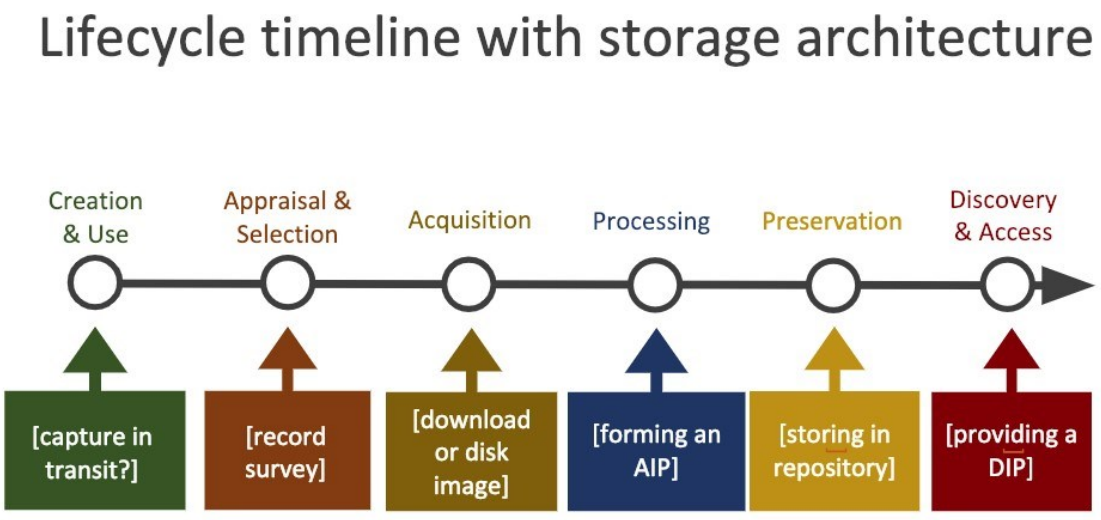


*Figure 4: Email Processing Across the Lifecycle Courtesy Tricia Patterson, Harvard University Library*

In principle, this workflow differs little from those for other types of archival records since it aims at forming up and Archival Information Packet (AIP), managing it in preservation storage environment and providing

Dissemination Information Packets (DIPs) to end users. It can be implemented in any repository context, provided staff can secure access to the types of tools listed above and to a sufficient and, ideally replicated, storage solution. That said, most email collections will require multiple processing steps, and it is likely that some steps, such as appraisal or processing, will require additional refinements through iteration.

The report of the Task Force on Technical Approaches for Email Archives provides sample processing workflows, incorporating either migration or emulation (Task Force on Technical Approaches for Email Archives, 2018b, pp.68–76). These workflows were developed by the Task Force members, who tended to represent email innovators and larger academic institutions. They should be studied carefully, but not all repositories will be able to chain together tools into the types of sophisticated workflow that they represent. Such institutions should start small, working with the basic bit-level processing workflow.

Figure 5 provides a slightly more complex version of such a bit-level workflow than that provided in *The Future of Email Archives* report, but one which should be achievable in most repository contexts. This workflow envisions email from acquisition to point of access. Originating from one of several accounts, email is processed through commonly available process tools. At the end of the process, metadata is created in a catalogue or archival management system, then an MBOX file (with attachment written out as binary content and referenced in the messages) is deposited in a preservation repository serving as a dark archive. Users query the catalogue and request access to the collection as a data packet, which they can use with their own software.

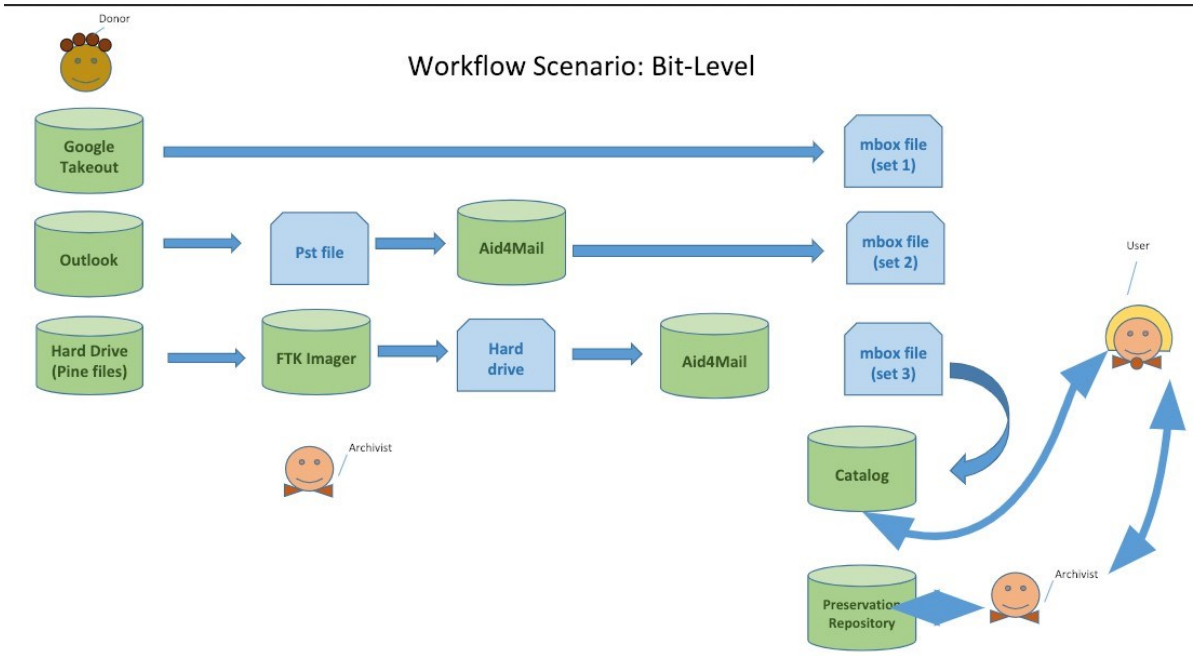


*Figure 5: Expanded Bit-level Preservation Workflow (Adapted from Task Force on Technical Approaches for Email Archives, 2018b, CC-BY-NC-SA)*

This workflow should not be idealized. Repositories can add or subtract processing actions and software tools as needed, at several steps along the way. For example, simple virus checking could be added on the MBOX file once it is extracted. ePADD or another tool could be used to appraise the MBOX file, and to remove sensitive content, as well as to generate metadata. DArcMail could be used to package the MBOX file into EAXS format for preservation, while the MBOX file itself is treated as a dissemination packet.

Over the next few years it is likely that workflows will undergo additional automation and smoothing, as the recommendations from the Email Archives Task Force are implemented (Task Force on Technical Approaches for Email Archives, 2018b, pp.76–89). Yet the promise of future improvements should not excuse t repositories from the duty to make forward progress in the present. The steps outlined above can be pursued with even the simplest of equipment and tools.

## 7. Case Studies

Although the preceding sections provide useful information regarding tools and workflows, the following case studies directly illustrate the range of issues that repositories face in preserving email. While no single solution will apply across the board, each case illustrates that planning, relationship building, and technology selection play equally important roles in facilitating success.

### 7.1. Developing an Email Policy

*This case study is based on an interview with Christina Somovilla, Records and Information Governance Manager for the Financial Ombudsman's Service, UK.*

The United Kingdom's Financial Ombudsman Service is an independent public body, formed by Parliament in 2000 and comprising about 4,000 public-sector employees. It acts as an independent arbiter to investigate and resolve disputes between consumers and financial service organizations, holding the power to settle consumer complaints against banks, insurance companies, payment reconciliation services, pensions, and other companies (Financial Ombudsman Service, n.d.). Its decisions are based on the law and a prepared view of facts. In the 2017/18 fiscal year, the Ombudsman Service received 339,967 new complaints and resolved 400,658 (Financial Ombudsman Service, 2018). Over the past five years, the number of handled complaints has doubled, largely but not exclusively in response to a well-reported financial scandal, the misselling of Payment Protection Insurance (Kollewe, 2017). The Ombudsman's rapid growth has led to an information proliferation, but the size of its records management staff has held steady at three FTE employees.

In an interview, Christina Somovilla, the records and information governance manager, stated that any information communicated or recorded within the organization is considered to be a record of business activity. As such, emails are subject to classification, disposal, and retention requirements. Like many staff at public bodies and government organizations, the records management staff at the Ombudsman Service have developed records schedules and disposal advice that reflect specific business activities or functions. Thus, emails can and do subsist (from a records perspective) within many record series, although they are sent and received using an email-specific platform, Microsoft Exchange.

Given the rising workload, senior management had historically shown rather passive support for records and information governance policies. Retention/records management was not seen as a priority and as a whole there was less appetite for information governance. While the Ombudsman Service complied with legal requirements, most line employees and managers – understandably – showed less appreciation for the potential historical or evidential long-term value of records, and the preservation of such emails was far from assured. The situation changed in 2016, when the European Union brought forward the General Data Protection Regulation (GDPR). GDPR increased the visibility of information governance issues, and tightened requirements on processing personal data. Records management staff used this opportunity to push email retention up the agenda. A recently established information governance board was used as a vehicle for the records management team to highlight both the risks of keeping unnecessary records and those of inadvertently destroying ones of long-term value.

Of all the record series that hold email, the case file is the largest in terms of extent and this is also the core work completed by the Service. Some cases do not move to investigation, and those records are deleted after three years. For cases that lead to an investigation, a view of the facts is generated. The view and supporting documentation must be retained for six years from the point the case closes. If the view is challenged by the

consumer or business, the case is assigned to an ombudsman, who has authority to issue a final decision. If the case goes to an ombudsman, the final decision is retained permanently, but supporting materials are destroyed six years from the date of decision.

Other functions also generate emails requiring long-term retention. These include governance records, such as board meeting minutes and supporting materials, records relating to executive decisions (and the broader decision-making process) and legal advice, which are subject to a ten-year retention period. But many records generated by transaction-based business activities, such as payroll records, recruitment records, or meeting scheduling, require much shorter retention.

Within this environment, only a small percentage of the emails generated hold enough long-term administrative, financial, legal, or historical value that they should be transferred out of the email system in which they were sent or received. Records Management staff would like to treat such emails according to the function or activity that generated each message, since the classification of email by functional record group would allow for the timely deletion of email and other related records with a purely transactional value, in many cases directly from the email server. Moving records with long-term value, such as those shedding light on policies and programs, into the correct shared administrative folder will allow the Service to preserve them. These records hold high value as the Service responds to new financial abuses, social pressures, and political changes.

Ultimately, the information governance board's advocacy efforts and work by the records management staff resulted in the development of an email and retention policy, which includes the following provisions:

- All email is considered an asset of the organization to be managed as such. It is not the property of the account holder, but of the Service.
- Its retention or disposition will be subject to records advice for a particular record series.
- Staff members should file email related to a particular complaint or case with the case itself or, for non-case emails, by cutting and pasting or moving the messages in the shared folder, on a network drive, or other appropriate system where retention rules can be applied. Decisions, which are issued as PDF documents, will be stored separately, facilitating their long-term preservation, while associate documents such as email are automatically deleted at the end of the retention period.
- Shared inboxes will be subject to defined retention periods. Most shared inboxes, such as those to support particular cases and no longer in use, will be eliminated. For all new case files, emails will be retained for six months before automatic deletion.
- A few shared inboxes, such as those supporting legal work or subject access request, will be granted much longer retention.

In summer 2017, the information governance board provided critical support in drafting and securing senior management agreement to the policy. Staff subsequently shared it with select stakeholders, including data and record owners, support staff, and senior members of the leadership team. In early 2018, the policy was formally shared Service-wide on the staff Intranet, and an FAQ developed. It generated a few queries and was fully implemented in September 2018.

To monitor and assess the new approach, records management and IT staff monitor the shared network drive and personal drives for email, so see if it is being moved into that location. They also generate reports to see how much email exists, where it is growing, shrinking, or otherwise changing.

## 7.2. Decommissioning Lotus Notes and Improving Email Governance

*This case study is based on an interview with Bridget Sisk, Chief of UN Archives and Records Management*

*Section (ARMS), Monika Tkacova, Information Management Officer, Baha Al-Attia, Information Systems Officer, and Hasib Beg, Business Analyst. Additional information was incorporated from a private internal report regarding a database appraisal project (Tkacova, 2018).*

The United Nations faces records management and archival challenges that dwarf those faced in many other organizations. Headquartered in New York City, the UN includes many constituent agencies offices, and missions, composed generally of central operations and of the peacekeeping, humanitarian, and development operations which themselves make up over 76% of its annual budget ($48.8 billion USD in 2016) (United Nations System Chief Executives Board for Coordination, 2016). This case study focuses on the records of the peacekeeping operations, and specifically on the appraisal of databases in a Lotus Notes system that supported their operations. While the appraised records were ultimately not found to include email, the project still provided useful lessons that can be applied to email appraisal and assessment.

Since its inception, the UN has generated correspondence that warrants long-term preservation, including traditional, paper-based, correspondence, fax transmissions, encrypted communication from field missions to UN Headquarters (code cables), and now electronic communications (email records). Historically the UN Archives and Records Management Services (ARMS) has provided access to a great deal of correspondence, including in some cases digitized code cables, which are included with the records of the Secretaries-General Fonds (record group or classification). In addition, all code cables are considered a permanently valuable record and are entered into two separate repositories: one maintained by the UN's Office of Information Communication Technology includes all code cables sent or received by UN Headquarters since 2007, the second one includes cables that were only addressed to the Department of Peacekeeping and the Department of Field Support (2012–2018). In January 2019 a new repository was implemented to provide one-stop-shop access to all code cables. Code cables in all repositories are isolated from other recordkeeping systems, and are kept in a standardized format, which preserves their value as evidence. The specific messages that have been sent, received, and distributed document decision-making processes and the activities of missions.

Alongside the code cables, UN staff have implemented email systems to meet specific headquarters and field missions' needs. After experimenting with cc:Mail in the late 1980s and early 90s, the UN began employing client/server email systems around the year 1995. From a records management perspective, emails were viewed as transitory; official advice from the year 2000 specified that email should be automatically discarded after 30 or 90 days, with record copies of messages warranting preservation under existing records schedules filed in appropriate administrative systems prior to deletion (United Nations, 2000, pp.23–24). Shortly thereafter, and in response to an audit of the Oil for Food Program, the policy was revised to specify that email messages should not automatically be deleted and that accounts should be retained. As important as this shift was at the time, such information governance advice has been difficult to implement, given the UN's international structure and dispersed information management model.

Accordingly, Archives and Records Management Section staff seek opportunities to prove the value of good records management. Ideally, it allows for the disposal of transactional records, while preserving those that shed substantive light on the UN's activities, achievements, and controversies.

While all UN staff send and receive code cables using a common system, the peacekeeping missions have, historically, run their own IT operations, developing the specific functionality that they need to run the mission. From 1995 through 2018, the Office of Information and Communication Technology Directorate (OICT) provided each mission with access to Lotus Notes/Domino, which served as a rapid development environment (Wikipedia, 2018b). Using the workflow functionality provided by Domino, IT staff members developed bespoke applications supporting the precise asset management, human resources, purchasing, transportation and communications functions that each mission required. The system's flexibility and customizability were critical virtues, as both the number and complexity of peacekeeping operations has expanded rapidly over a 25-year period. Yet its use complicated the application of information governance

principles. Thus, historically little control was exercised over the design of the IT systems used by the peacekeeping mission, much less the records encoded within them.

In 2014, ARMS and OITC drafted an information strategy. While the strategy's information governance advice was never fully implemented, the UN began to migrate records and to consolidate IT platforms in FileNet, and to an even greater extent in Sharepoint. These decisions provided staff members in ARMS and OITC with an opportunity to assess Notes databases, where they hoped to discover inactive databases and begin to plan for their disposal or preservation. The project therefore provided opportunities to advocate for improved governance over database records.

At first, improved information governance for email did not seem like an obvious, or even potential, outcome. This is because all the Lotus records appraised as part of the project were actually transactional data, not email. The peacekeeping missions used Notes/Domino to automate routine processes, facilitate communication and collaboration, and ensure other operational needs. Most of the remaining data were the results of transactional processes, documenting actions as consequential as investigating human rights abuses or as mundane as paying routine invoices.

Each of these actions left behind a trail of records. Data was produced in the client application (Notes) and stored in the servers (Domino). Domino uses No-SQL-like data format, encoding all data in Notes Storage Format, or NSF, files (Sustainability of Digital Formats, 2015; Metz, n.d.).

As part of the assessment project, OITC staff identified over 6,000 individual databases stored in NSF files. The vast majority of these did not contain email data, but each database was replicated by OITC staff to a centralized UN Data Centre. A subset, about 1,000 of these databases, underwent analysis by archives and records management staff members. Many of the databases had been generated by a mission that had closed or completed its operations, and the data was of an operational and transactional nature. For this reason, ARMS and OITC staff suspected that many of the databases could be discarded or migrated to another system. A much smaller percentage likely held information of long-term, archival value.

To demonstrate the value of information governance and archival appraisal, ARMS and OITC staff decided to focus first on the assessment of transactional data, leaving the email stores located in the NSF files for a subsequent phase. One Information Systems Officer imported metadata about the NSF databases to a spreadsheet, and an Information Manager appraised 1,006 of them, grouping them by function and subfunction within each mission.

Using the file name as a key to the content, Tkacova browsed the databases with Notes web tools, opened selected databases, examined their content, and mapped them to business processes. Tracking them with a spreadsheet, she noted that 950 of the databases contained transactional data that could have been previously discarded under existing records schedules (United Nations Archives and Records Management Services, n.d.). Fifty of the remaining databases required some additional retention, but showed no permanent value, and only two were of clear archival value warranting long-term preservation. These contained policies and were recommended for transfer to the archives. Based on these recommendations, the other data has been recommended for deletion. If the individual offices that own the data wish to retain it, the costs for data centre storage will be charged back to them.

Overall, the project has been successful in demonstrating the value of appraisal of email by function and alongside related transactional data, leading to a significant potential cost saving in reduced software licensing and storage costs. A large volume of records was assessed after their active use had concluded, leading to the assessment of email stores alongside the entire record of business processes. The project has also generated criteria that can be applied to the remaining 5,000 databases, facilitating subsequent appraisal processes.

As a next step, ARMS staff plan to appraise specific email collections and to make a case for their preservation. From a technical point of view, the preservation of email from the missions is seen as a relatively straightforward matter — migrating the files to PST format if the decision is made to retain the files. By taking a small-scale, thematic look at accounts from a historically significant mission, ARMS staff hope to positively inform conversations within the UN, leading to the preservation of the small number of substantive records shedding light on the results of peacekeeping activities, while also allowing for the disposal of other records, at significant saving to the organization.

## 7.3. Processing Capstone Email Using Predictive Coding

*This case study was written by Brent M. West and Joanne Kaczmarek, University of Illinois at UrbanaChampaign.*

The University of Illinois is collaborating with the Illinois State Archives on a project to preserve and make accessible email of former officials working in the Office of the Governor of the State of Illinois. The Governor is responsible for the execution of the laws in Illinois and for numerous agencies, including the Department of Innovation & Technology, which manages IT systems for most offices under the Governor. The Archives, a department under the elected Secretary of State, administers the State's records management program and is the depository for records judged to have long-term value in documenting the activities of state agencies.

The Archives and Office of the Governor both wish to responsibly manage State records for the benefit of its citizens, but their priorities differ. While the Archives is entrusted with the preservation of the State's history and its records to support transparency, accountability, and citizen's rights, the Office of the Governor and its agencies create, use, and maintain records in support of efficient operations of State government, execution of laws, and policy initiatives. The duties, politics, and turnover within any public office can influence records management practices to inhibit public access, impede collaboration, or limit documentation of politically sensitive information.

The email of executives is often considered the modern equivalent of correspondence files, long held to have enduring value for office operations and for researchers. Following the Capstone approach described earlier in this report, the Archives appraised records from the previous two Democratic administrations. (The current Governor is Republican.) The Archives identified senior officials of the Office of the Governor and its agencies whose email should be preserved. The email has been held by the Department of Innovation & Technology in Enterprise Vault since 2008, along with the email of nearly all state employees under the Governor.

Email poses unique challenges to archivists, however. It is difficult to acquire and process due to its quantity, diverse topics and formats, mix of personal and official communications, inconsistent organization, and sensitive content. Open records laws and account size limits can result in overzealous purging. Administration, leader, and staff turnover, as well as operational needs and limited IT staff experience preserving historic records, can lead to undesirable account deletions.

In 2015, the Archives secured permission to acquire the email of the former officials by working with the Governor's legal counsel to address the records management needs of the new administration. As time allowed over the next several months, the Department of Innovation & Technology exported the data (500 GB/5M messages) in PST format from Enterprise Vault onto an external hard drive. The data was then copied to Preservica and a secure network share at the University of Illinois for processing. Although additional accounts should be preserved, and thousands of accounts of former state employees could be deleted, a two-year budget impasse, high workload, and the upcoming election have deferred any additional transfers or reviews for deletion.

To address the challenges of email, our project is evaluating predictive coding technologies developed for the legal community. Predictive coding has the potential to alleviate the challenges of processing email by using supervised machine learning to augment appraisal decisions, identify and prioritize sensitive content for

review and redaction, and generate themes and trends. We have evaluated software including: Luminoso, Office 365 Advanced eDiscovery, Recommind, Ringtail, and TAR Evaluation Toolkit. In addition, we have explored related tools including AutoTAR, BitCurator Bulk Extractor, ePADD, and TOMES. To address challenges posed by the proprietary PST format, Emailchemy will be used to transform the source files to MBOX format and to generate unique identifiers for the email messages and attachments.

A predictive coding model is trained by manually coding several hundred documents sampled from the collection. Algorithms use these coding decisions and attributes of the documents such as word distance and frequency to rank each of the documents in the collection on a scale from -1 to 1, meaning the document is unrelated or related to the inquiry (e.g., is the document archival?). Many of the tools are capable of importing PST files, extracting technical metadata, text recognition, non-textual document characterization, deduplication, and complex multifaceted searches.

Once processed, the original PST files will be replaced in Preservica by the subset which has been identified as archival, including technical and descriptive metadata generated by the tools and processing details, and attachments will be assessed for format migration needs. The non-archival subset will be retained temporarily for several years as a precautionary measure while the workflow matures. The non-restricted archival content will be made available initially with Microsoft Outlook on a secure terminal in the Archives reading room, while only header metadata will be provided for restricted content. Restricted content will be redacted on demand.

Early results are promising for the potential value that predictive coding can provide to archival workflows. For a subset of 200,000 messages, for example, 2,000 messages were manually coded as archival or nonarchival and 6,900 were manually coded as restricted or public to achieve 81% and 66% recall, respectively, based on the models' performance against a control group. For comparison, studies indicate that manual human review of an entire corpus typically achieves less than 65% recall. The variance in effort and success appears largely due to differences in prevalence: 42% of the documents are archival versus 2.6% restricted. Most restricted documents were routine personnel matters which are more easily weeded out as nonarchival. Keyword libraries and regular expression searches can supplement this review for obvious sensitive content.

Human factors remain a challenge. Coding a document sounds simple, but the results can be affected by differences in reviewer knowledge of the subject matter, coding experience, and fatigue. Retaining entire accounts permanently as correspondence files should be reconsidered if 40 to 60% of it is non-archival or outright spam. Acquisition remains a concern because only 68 email accounts were located for the 181 officials identified by the Archives. Complex or non-textual email attachments such as images, multimedia, binaries, databases, and encrypted files will require separate processing. While access will initially be limited, options to make the content available online and/or through some of the tools mentioned above is desirable. Cost may be a factor for some institutions, although some low-cost, open-source, or consortium options may be sufficient.

While much work remains for our email preservation project, the progress in recent years to stem the losses and develop a sustainable path towards access is an achievement. Our hope is that this work empowers records creators, archivists, and researchers to better understand, synthesize, protect, and preserve email collections.

## 8. Conclusions and Recommended Actions

While finishing this report, I read about high school teacher Chuck Yarborough. In his history classroom, students complete *Tales from the Crypt* projects. Coupling archival research with performance art, they develop a deep and nuanced understanding of social relations and racism in Mississippi during and after the

US Civil War (Rizga, 2018). People like Yarborough and his students remind us that archives provide individuals and society with a unique power: they connect the past to the present. People can use archives as a form of time travel – they can look back and, hopefully, build present-day understanding and empathy. And in the best of all worlds, that leads to actions that build better tomorrows.

If we wish to provide future generations of teachers and students with these powers, present-day archivists can, and must, take some simple steps to preserve email. Specifically, the following actions seem particularly warranted at this time.

## 8.1. Recommended Actions for Individuals

Anyone can take basic preservation steps with his or her own email and can help others do likewise. It all starts with understanding your own email account and in seeing that email has a value that outlives its immediate use. In the case of work email, the value may seem obvious: who amongst us has not had to track down a key report, finding or recommendation? Might your colleagues or successors need access to your emails, in order to understand institutional history, policy-making or decision-making?

And what about personal email accounts? Your messages may have sentimental value to you or family members. If you are the type of person whose activities shape or reflect social, political or cultural life, your account may hold historical value, a bit like the handwritten or typescript letters from yesterday. For these, and many other reasons, you should take step to preserve your emails, and possibly even to prepare it for transfer to an archives.

What's next, once you've decided to preserve your emails? Take a bit of time to understand how the technologies you use store and manage messages. Read local documentation about storage limits and deletion policies. Reach out to IT staff to better understand any management features, such as retention policies and archiving features. Understand the difference between the email servers (which send, receive and store the 'authoritative' copy of your messages) and the client programs (which provide a gateway to the server, and may also store a copy of the messages). By uncovering a few technical facts, you can make sure that you are not inadvertently losing or misplacing messages.

Next, it is time to implement some personal best practices. There is no one-size-fits-all solution, but a menu of potential actions, such as those shown in the boxed content below, may help you to put a framework into place:

***Personal Email Management Practices***

- Know where and how your email messages are stored.
- Delete spam and non-actionable messages (e.g., listserv notifications) as they are received.
- Ensure that email stores are synced across devices (i.e., that sent messages are saved to the same server location).
- Configure email clients to keep a remote copy of all sent messages.
- Segregate old messages from your inbox to a separate 'archive' or 'non-current' folder, where you are less likely to delete it.
- If emails have not been filed as received and sent, use filters to develop project or topic-specific folders.
- If possible, store all messages on the server, not in local storage locations.
- Be aware of server storage limits.
- Understand both the capabilities and limits of email search and folder systems.

- Undertake bulk deletion and transfer actions with great caution.
- If server storage is unlimited, or practically so, use it as a medium-term archive.

- If messages must be removed from the server, export and store messages in MBOX, PST, or another format and location where they can be browsed using your standard email access tools. If your organization does not provide a centralized space for such files, it may be necessary to work with ITC staff to identify one.
- Take some simple steps to preserve your emails. Use cloud-based email services such as Gmail or Outlook365 to download and preserve email, and find out how to better manage email in your existing systems.
- PRO TIP #1: Explore the use of personal email archives tools to create your personal email archives.
- PRO TIP #2: Long-term email users may wish to identify old email stores left on computers and devices, then centralize them in one location, and use tools such as those identified in this report to convert them to a preservation-ready format, such as EML, MBOX, or (less ideally) the proprietary but open and commonly used PST format.

Ideally, you should consult closely with a records manager, archivist, or an ICT staff member for decisionmaking help. They can help you implement the best practices from among a range of choices. Remember that actions that work in one organization may fail in another. For example, if your system gives you liberal storage limits, your best medium-term preservation option may be to do nothing: you can let the messages accumulate on the server, where they may be preserved under institutional policy. (Thankfully, built-in storage limits are becoming less and less common, as more and more organizations implement Infrastructure as a Service email options, such as Microsoft's Outlook 365 and Google's Gmail service, which is part of their G Suite service.) But caution is warranted, since some organizations configure servers with hard limits or time-based deletion policies, in line with records management advice. In these cases, it is still possible that copies of messages are being captured and stored separately, using an email archiving application. So, the best practice is always to understand institutional policy and procedure, then adapt your own actions to suit that framework.

The types of actions listed above will help you preserve your messages for your own use and perhaps to better prepare them for long-term preservation. Keep talking to archivists or records managers. You'll soon learn whether your files might be suitable for long-term preservation in a formal archives program. There is no need to immediately open up your entire email record to public inspection: archivists will help you shape a policy that accords with institutional requirements and personal wishes. They can help remove materials that might be irrelevant, shaping the collection to reflect your wishes.

## 8.2. Recommended Actions for Institutions

You've made the decision. Your archives would like to help staff members or donors preserve email. What's next? There are three basic steps which institutions undertaking email preservation projects should take: defining policies, choosing appropriate tools, and implementing them in the light of local environmental factors and available resources.

***Policy Setting:*** Email policies should outline an institutional commitment to email preservation and list specific actions that will be taken to implement it. Policies must define categories of email that are critically important for administration, institutional memory and cultural value. Procedures will define how systems support the policy and how users interact with systems, allowing an organization to manage email over its entire lifecycle. User policies should be concise, laying out acceptable uses and the need to comply with company policies regarding privacy, data protection and records law (Smallwood, 2008, pp. 21–39). Most importantly, they should facilitate good personal information management principles by providing reasonable amounts of storage or, preferably, recourse to an email archiving tool. Users benefit from a technical framework that automatically segregates archived messages to an external, server-based repository, so such a method will often play a key role in the preservation strategy. In many institutions, a 'Capstone' policy, which attempts to identify key decision-makers and preserve all messages from their accounts, should be strongly considered.

Another key policy will outline end-user expectations, responsibilities, and rights regarding the access, use, privacy, and control of email archives. Such policies will not only benefit the end users but will also engender trust in those from whom email is being acquired. Naturally, many people have concerns about their emails or other data being placed into a repository or access system, even if the system is a dark archive, where access is mediated by an archivist or records manager. When an archive explicitly defines the ways in which staff will weed, process, or otherwise appraise the email, and when the institution spells out access methods, restriction periods, and acceptable uses of the preserved accounts, staff are much more likely to support the policy. This is particularly true if non-archives staff are consulted or given a strong voice in defining the policy, whether they include legal counsel, top administration, or line staff members.

***Choosing Tools:*** After defining policies, institutions should experiment with select tools. As noted above, there are many to choose from, and the *Email Archives Task Force Report* includes helpful tool lists and descriptions, valid at the time of writing (Task Force on Technical Approaches for Email Archives, 2018b).

The list of available tools will change over time, but repository staff should evaluate tools in the following functional areas, which are less likely to change over time:

- *Appraisal and Capture:* First, you'll need to decide what to capture: the entire account, or only portions such as the sent or received messages, or a specific set of folders. While repository policies should guide the archivist, the decision should also be based on an analysis of the account. Once you've decided *what* to capture, there are several technical options to consider for *how* to capture messages for additional processing. These include 'email archiving' tools, which capture each message as it is sent or received, server-based export tools (such as those built into email clients applications or webapps), forensic or disk imaging technologies (which make a bit-for-bit copy of local message stores), and email-specific migrators (which can import messages or transfer them from one format to another). Unless your repository deals only with emails coming from one source system, it is likely you will need to mix and match capture tools, based on the source you are dealing with.
- *Stabilization:* Once email has been moved from its source to a holding area, archives staff may need to convert it to a common, preservation-ready format, from which state additional processing actions and steps can be taken. In some cases, it will already be in one of these formats, such as MBOX or EML. But if not, staff can use paid or open-source tools to convert the message. Optionally, attachments can be converted to binary files, then processed and stored separately.
- *Triage and Removal:* A range of possible options present themselves here, including relatively expensive enterprise-level programs for classification and legal discovery. Open-source software will likely be more immediately accessible to archives staff. Questions such as the following should be carefully considered when using the tools:
    - What methods does the tool use to sort and classify messages?
    - How can staff mark or exclude certain messages or types of message for removal?
    - Does the tool provide a log or actions record?
    - What technical methods does it employ (such as natural language processing, or machine learning)?
    - How is the human user kept in the decision-making loop?
    - Does the tool supply metadata that can be used in subsequent stages of the processing workflow?
    - Will it export messages in a preservation-friendly format?
- *Arrangement and Description:* It may seem that arrangement is unnecessary when dealing with email. Aren't accounts already in an original order, and shouldn't they be kept that way? Yes, but tools that classify and extract metadata from them can present alterative arrangements or access pathways through the collection, such as by an extracted entity (a place or personal name or a term drawn from a subject-specific lexicon). These tools may also supply some exportable metadata,

either at the message or collection level. Item-level metadata, however, should not be seen as a substitute for a good series or collection-level description, which can be developed by an archivist using the repository's standard collections management software, under the terms of national guidelines and local policies for description.

- *Storage and Access*: In many cases, email archives will be loaded into preservation management systems that store multiple types of data, not just email. For these reasons, repository staff should take considerable care to shape the archival packet in a way that it can be easily ingested into that system, while respecting the collection's provenance and order. A critical decision concerns attachments. To make them preservation-ready and migrate-able objects, it will be desirable to have attachments decoded from MIME to their source binary file. However, doing so runs the risk that the attachments may become decoupled from the parent message, particularly if content is moved from one system to another, at some future point in time. For these reasons, it is advisable to keep a copy of the original source account in PST or MBOX. This may not be necessary, if the repository service has a very active management plan for attached content.

***Implement Workflows:*** Ideally, the tools that you choose will be chained into replicable procedures; that is to say, workflows that can be applied across many email collections. In practice, this means that institutions will need to define an email accessioning and processing workflow. As part of this packet, the message content (the messages and attachments) should be preserved in a common format across collection types. Messages may originate from many different systems or accounts, so an important early step will be converting them to a standardized format, such as MBOX. From that point, a more or less linear workflow can be developed. Optionally, this will include a step to winnow the corpus down, removing non-record emails or those that must be removed under a donor agreement or recommendation.

A key outcome from this process will be the formation of an archival information packet structure for the email account, so that messages can be preserved in a format that will allow for deposit into the archives' existing preservation repository. There is no need to reinvent the wheel. The structure may be as simple as a folder containing the messages in PST, MBOX, or EML format, along with a separate copy of the attachments. All of this can be ingested to the standard preservation repository. In addition, collection or series-level metadata should be produced and included within an archives' collection management or finding aid applications. These actions will provide a foundation of email preservation, upon which an archive can build additional features, processing, actions, and access methods.

Like other areas of practice, email preservation is not a static field. As new tools and methods are developed, an archives' staff member can assess them and, if warranted, add them to an email preservation and processing toolkit. For example, repository staff may experiment with tools that support machine classification or natural language processing of collections. Such tools are being included in applications such ePADD and TOMES. These tools can be used for several purposes, including record appraisal, weeding, classification, description, and discovery.

## 8.3. Recommended Actions for the Community

Email preservation is now a more common part of the preservation workflow, and repositories can chain together tools to preserve messages and make them more accessible to end users. But email preservation is still an endeavor that is more complex and less settled than other areas of digital preservation practice. Those interested in email preservation will find many opportunities for collaboration with colleagues in other institutions. For example, the community might support initiatives around the following topics:

- Supporting existing software projects developed for the appraisal and processing of email.
- The development of a PDF/A-based standard for email, which could be used to encode messages and attached comments within a preservation-friendly format.
- Methods to capture content that is linked to messages via embedded urls.

- Self-archiving applications, which people could use to create a mirror of their own accounts, then pass it along to colleagues or gift it to a repository.
- The development of open-source 'email archiving' services, which might capture and file messages at point of transmission, rather than after the fact, much like social media archiving applications.
- The development of additional tools to apply natural language processing, predictive coding, and machine learning to the problems of email appraisal and classification.

## 8.4. Conclusion

In the 2011 edition of this Report, I cited several ripped-from-the-headlines incidents to demonstrate both the interest that email occasions among the public and also its fragility. Far from declining in importance, email continues to fascinate people and play a prominent role in the news of the day or journalistic exposés. Behind these incidents, archives can and must work to ensure that future generations have access to the sources on which these first drafts of history are being written.

Thankfully, the right tools are now at our fingertips. It is up to us to use them wisely. At the same time, archivists and other information professional should continue advocating for the important role that email archives can play not just in documenting activities, but also in fostering values such as professionalism, collectivity, activism, selection, preservation, democracy, service, diversity, use, access, and history (Greene, 2009).

## 9. Glossary

**Application Programming Interface (API)**: A set of routines, protocols, and tools designed for developers to build applications on top of the underlying building blocks of an original piece of software, allowing the underlying software to hide particular implementations while still sharing the information needed to create new applications.

**Domino Server**: A proprietary application developed originally by the Lotus Corporation, and now owned, developed, maintained and licensed by IBM. Domino provides an email server/message transfer agent and several other features, including calendaring, scheduling, and task management. Domino servers are typically used in tandem with the Lotus Notes client/user agent. They are known for their replication features, which allow system developers to easily make synchronized copies of data on another server, or on the local user's desktop computer or on another device (IBM Corporation, 2009). Depending on system configuration, users may be able to connect to a specific Domino server using any IMAP-aware client application.

**ePadd**: A software package developed by Stanford University Special Collections & University Archives that supports archival processes around the appraisal, ingest, processing, discovery, and delivery of email archives.

**Exchange Server**: A proprietary application, developed and licensed by Microsoft Corporation, providing server-based email, calendar, contact, and task features. Exchange servers are typically used in conjunction with Microsoft Outlook or the Outlook Express web agent. They use a proprietary storage format. Messages sent using Exchange usually include extensive changes to the header of the file, and calendars, contacts and tasks are also managed via extensions to the email storage packet. Depending on local system configuration, users may be able to connect to a specific Exchange server using any IMAP-aware client application.

**Internet Message Format (IMF)**: A defined syntax specifying the precise set of rules by which a text file may be sent between computers as part of an email system. Defined most recently in the IETF's RFC 5322, IMF does not provide for the transmission of non-text-based files, such as binary application files, images or attachments. Rules for including those files are included in the suite of protocols defining Multipurpose Internet Mail Extensions (MIME).

**Internet Message Access Protocol (IMAP)**: A code of procedures and behaviours regulating one method by which email user agents may connect with email servers and message transfer agents, allowing an individual to view, create, transfer, manage and delete messages. Typically contrasted with the POP3 protocol, IMAP is defined in the IETF's RFC 3501. Email clients connecting to a server using IMAP usually leave a copy of the message on the server, unless the user explicitly deletes a message or has configured the client software with rules that automatically delete messages meeting defined criteria.

**Internet Engineering Task Force (IETF)**: An informal, open group of system engineers, vendors, computer operators and interested individuals who define the standard protocols by which the Internet operates, via a set of working groups and meetings. The IETF issues Internet standards in a Request for Comments (RFC) format.

**MAPI**: see **Messaging Application Programming Interface**

**Messaging Application Programming Interface (MAPI)**: A proprietary but open protocol for accessing and manipulating messages stored in the Microsoft Exchange Server and related parts of the Exchange/Outlook architecture on a Microsoft Windows computer. By defining a set of objects, functions, and methods, Simple and Extended MAPI can be used to add messaging functionality (including message creation, transfer, deletion and categorization) or to develop applications to capture and store email from an Exchange server.

**Message Transfer Agent (MTA)**: Software that transfers a message from one computer to another within a client–server architecture defined by the Simple Mail Transfer Protocol. Multiple MTAs may handle a message before it is delivered to its final destination.

**Migration**: The process of converting an email message or messages from one storage format to another storage format. Migration can be completed using tools built into an MTA or UA, or by stand-alone migration tools, such as Xena, Aid4Mail, and Emailchemy.

**MIME**: see **Multipurpose Internet Mail Extensions**

**Multipurpose Internet Mail Extensions (MIME)**: A protocol for including non-ASCII information in email messages. Specified in IETF RFC 2045, 2046, 2047, 4288, 4289 and 2049, MIME defines the precise method by which non-Latin characters, multipart bodies, attachments, and inline images may be included in email messages. MIME is necessary because email supports only seven-bit, not eight-bit ASCII characters. It is also used in other communication exchange mechanisms, such as HTTP. Software such as message transfer agents, email clients, and web browsers typically include interpreters that convert MIME content to and from its native format, as necessary.

**Post Office Protocol (POP3)**: An Internet protocol that defines the ways in which an email user agent may connect to an email server to retrieve and manage email messages that the server or client is holding in storage. POP3 typically moves email messages from the server to the client machine and deletes the server copy, although it is possible to configure the server to maintain the message, or order the server to do so, via a setting in the client application.

**PST**: .pst is a file extension for local 'personal stores' written by the program Microsoft Outlook. PST files contain email messages and calendar entries using a proprietary but open format, and they may be found on local or networked drives of email end users. Several tools can read and migrate PST files to other formats.

**Simple Mail Transfer Protocol (SMTP)**: A set of rules that defines how outgoing email messages are transmitted from one Mail Transfer Agent to another across the Internet, until they reach their final destination. Defined most recently in IETF RFC 5321.

**TOMES Tool**: Developed by the State Archives of North Carolina, with support from the State Archives of Utah and Kansas State Historical Society, this email preservation and processing tool allows archivists to process complete email accounts more quickly by using NLP tagging to identify PII, confidential information, and named entities using dictionaries specific to state government in an XML format.

**User Agent**: Software that interacts with an email server to retrieve and send messages, and with the end user to create, store, edit, delete, print, classify and otherwise manipulate email messages.

**Unstructured Data/Records**: Data or records that do not conform to a specified data model or which cannot be queried using a standardized syntax, but which are stored in a file system. Records such as email messages, correspondence stored in personal workspaces, text/instant messages, and blog postings tend to include unstructured data, although email headers provided some structured data for each message.

## 10. References

Archives and Records Association 2017, 'When Worlds Collide – Records Managers and the GDPR – Challenges and Opportunities?' ARA Blog (blog). June 1, 2017. https://www.archives.org.uk/news/blog/item/when-worlds-collide-records-managers-and-the-gdprchallenges-and-opportunities.html Last Accessed 02/05/2019.

Archivematica Wiki. 2017. "Email Preservation." March 23, 2017. https://wiki.archivematica.org/Email_preservation Last Accessed 02/05/2019

Ashenfelder, Mike. 2011. "Personal Archiving in the Cloud." *The Signal: Digital Preservation* (blog). June 9, 2011. http://blogs.loc.gov/digitalpreservation/2011/06/personal-archiving-in-the-cloud/. Last accessed 02/05/2019.

Baker, Fran 2014, 'Carcanet Press Email Preservation Project Phases 2–3: Final Report'. Carcanet Press Email Preservation Project. United Kingdom: Carcanet Press. https://www.escholar.manchester.ac.uk/api/datastream?publicationPid=uk-ac-manscw:226625&datastreamId=FULL-TEXT.PDF Last accessed 20/04/2019.

Baker, Fran 2015, 'E-Mails to an Editor: Safeguarding the Literary Correspondence of the Twenty-First Century at The University of Manchester Library.' *Special Collections in a Digital Age*, Vol. 21:2, 216–24. https://doi.org/10.1080/13614533.2015.1040925 Last accessed 20/04/2019.

Baron, Jason 2010, 'The Future of Email Preservation', presented at NARA RACO 2010, Washington, D.C., May 12. www.archives.gov/records-mgmt/pdf/baron-raco2010.pdf Last accessed 20/04/2019.

Bailey, Jefferson 2013, 'Disrespect Des Fonds: Rethinking Arrangement and Description in Born-Digital Archives.' *Archive Journal*, no. 3. http://www.archivejournal.net/issue/3/archives-remixed/disrespect-desfonds-rethinking-arrangement-and-description-in-born-digital-archives/ Last accessed 20/04/2019.

Bearman, David 1994, 'Managing Electronic Mail.' In *Electronic Evidence: Strategies for Managing Records in Contemporary Organizations*, Ed. David Bearman, pp.176–207. Pittsburgh: Archives and Museum Informatics.

Bearman, David 2017, 'Review of "Office of the Secretary: Evaluation of Email Records Management and Cybersecurity Requirements, ESP-16-03".' *The American Archivist*, 80:2, 459–62. https://doi.org/10.17723/0360-9081-80.2.459 Last accessed 20/04/2019.

Bernstein, Joseph 2017, 'Here's How Breitbart And Milo Smuggled Nazi and White Nationalist Ideas Into The Mainstream.' *BuzzFeed*. October 7, 2017. https://www.buzzfeed.com/josephbernstein/heres-how-breitbartand-milo-smuggled-white-nationalism Last accessed 20/04/2019.

'BitCurator Access' 2016, https://bitcurator.net/bitcurator-access/ Last accessed 20/04/2019.

Brodkin, Jon 2011, 'The MIME Guys: How Two Internet Gurus Changed E-Mail Forever.' *Network World*, February 2011. https://web.archive.org/web/20131102023405/http://www.networkworld.com/news/2011/020111-mimeinternet-email.html?page=1 Last Accessed 2/05/2019.

Brynjolfsson, Erik, Eggers, Felix, and Gannamaneni, Avinash 2018, 'Using Massive Online Choice Experiments to Measure Changes in Well-Being.' *Working Paper 24514*. National Bureau of Economic Research. https://doi.org/10.3386/w24514 Last accessed 20/04/2019.

Buckles, Greg. 2011. "Custodial Email Preservation – Email Infestation." *EDiscovery Journal*, no. March 2011 (March).

https://web.archive.org/web/20130106062146/http://ediscoveryjournal.com/2011/03/custodialemail-preservation-%E2%80%93-email-infestation/ Last Accessed 02/05/2019.

Bunn, Jenny, Brimble, Sara, Selene, Obolensky and Wood, Nicola 2015, 'Team Europe EU28 Project 2015–16: Perceptions of Born Digital Authenticity.' *InterPARES Report*. InterPARES Trust. https://interparestrust.org/assets/public/dissemination/EU28_20160718_UserPerceptionsOfAuthenticity_FinalReport.pdf Last accessed 20/04/2019.

Caputo, Linda, and Narva, Nivedetha 2016, 'Selecting an API or Technology for Developing Solutions for Outlook.' Microsoft Office Dev Center. October 20, 2016. https://docs.microsoft.com/en-us/office/clientdeveloper/outlook/selecting-an-api-or-technology-for-developing-solutions-for-outlook Last accessed 20/04/2019.

Carroll, Laura, Farr, Erika, Hornsby, Peter and Ranker, Ben 2011, 'A Comprehensive Approach to Born-Digital Archives.' *Archivaria* 72: 0, 61–92. https://archivaria.ca/index.php/archivaria/article/view/13360 Last accessed 20/04/2019.

Chapin, Larry, and Attfield, Simon 2018, 'The Reconstruction of Narrative in E-Discovery Investigations.' In Digital Preservation Coalition. https://www.dpconline.org/docs/miscellaneous/events/2018-events/1766dpc-email-ii-attfield-chapin/file Last accessed 20/04/2019.

Coburn, Alston. n.d., 'Adventures in Email Wrangling: TAMU-CC's ePADD Story – BloggERS!' BloggERS! | Society of American Archivists Electronic Records Section. https://saaers.wordpress.com/2017/09/05/adventures-in-email-wrangling-tamu-ccs-ePADD-story/ Last accessed 20/04/2019.

Cocciolo, Anthony 2016, 'Email as Cultural Heritage Resource: Appraisal Solutions from an Art Museum Context.' *Records Management Journal*, 26: 1, 68–82. http://dx.doi.org/10.1108/RMJ-04-2015-0014 Last accessed 20/04/2019.

Cormack, Gordon V., and Grossman, Maura R. 2014, 'Evaluation of Machine-Learning Protocols for Technology-Assisted Review in Electronic Discovery.' In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 153–62. ACM Press. https://doi.org/10.1145/2600428.2609601 Last accessed 20/04/2019.

Cormack, Gordon V., and Grossman, Maura R. 2017, 'Navigating Imprecision in Relevance Assessments on the Road to Total Recall: Roger and Me.' In SIGIR '17 *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 5–14. ACM Press. https://doi.org/10.1145/3077136.3080812 Last accessed 20/04/2019.

Cox, Richard J. 2008, *Personal Archives and a New Archival Calling: Readings, Reflections and Ruminations*. Duluth, Minn.: Litwin Books.

Crispin, M. 2003. "RFC 3501 - Internet Message Access Protocol - Version 4, Revision 1." March 2003. http://tools.ietf.org/html/rfc3501 Last Accessed 02/05/2019

Digital Preservation Coalition 2015, *Digital Preservation Coalition Handbook*. 2nd Edition. Digital Preservation Coalition. https://www.dpconline.org/handbook Last accessed 20/04/2019.

Digital Preservation Coalition 2017, 'Assessing Digital Preservation Readiness.' May 4, 2017.

https://www.dpconline.org/docs/miscellaneous/training/1677-assessing-readiness-getting-started/file Last accessed 20/04/2019.

Dollar, Charles M, and Ashley, Lori J 2014, 'Assessing Digital Preservation Capability Using a Maturity Model Process Improvement Approach.' https://static1.squarespace.com/static/52ebbb45e4b06f07f8bb62bd/t/53559340e4b058b6b2212d98/1398117184845/DPCMM+White+Paper_Revised+April+2014.pdf Last accessed 20/04/2019.

Enneking, Nancy 1998, 'Managing Email: Working Toward and Effective Solution.' *Records Management Quarterly* 32: 3, 24.

Ferriero, David S. 2016, 'Criteria for Managing Email Records in Compliance with the Managing Government Records Directive.' National Archives and Records Administration. https://www.archives.gov/files/recordsmgmt/email-management/2016-email-mgmt-success-criteria.pdf Last accessed 20/04/2019.

Financial Ombudsman Service 2018, *Annual Review 2017/2018*. http://www.financialombudsman.org.uk/publications/annual-review-2018/full-review.pdf Last accessed 20/04/2019.

Financial Ombudsman Service n.d., 'About the Financial Ombudsman Service.' http://www.financialombudsman.org.uk/about/index.html Last accessed 20/04/2019.

Foggo, Gavin, Gross, Susanne, Harrison, Brett and Rodriguez-Barrera, Jose Victor 2007, 'Comparing EDiscovery in the United States, Canada, the United Kingdom, and Mexico.' *Newsletter of the Committee on Commercial & Business Law Litigation, Section of Litigation, American Bar Association*, 8: 4. http://www.mcmillan.ca/Files/BHarrison_ComparingE-Discoveryintheunitedstates.pdf Last accessed 20/04/2019.

Garcia, Cardiff 2018, 'Internet a La Carte.' National Public Radio, Inc., May 31, 2018. https://www.npr.org/templates/transcript/transcript.php?storyId=615932894 Last accessed 20/04/2019.

Gibson, Jeremy 2018, Tomes-Docker. CSS. State Archives of North Carolina. https://github.com/StateArchivesOfNorthCarolina/tomes-docker. Last accessed 20/04/2019.

Goethals, Andrea, and Wendy Gogel. 2010. "Reshaping the Repository: The Challenge of Email Archiving." In *7th International Conference on Preservation of Digital Objects (IPRES2010)*. Vienna, Austria. http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/goethals-08.pdf Last accessed 02/05/2019.

Gollins, Tim 2009, 'Parsimonious Preservation: Preventing Pointless Processes.' In *Online Information Proceedings*, 75–78. The National Archives (UK). http://www.nationalarchives.gov.uk/documents/information-management/parsimonious-preservation.pdf Last accessed 20/04/2019.

Greene, Mark. 2009, 'The Power of Archives: Archivists' Values and Value in the Postmodern Age (with an Introduction by Dennis Meissner).' *The American Archivist*, 72: 1, 13–41. https://doi.org/10.17723/aarc.72.1.k0322x0p38v44l53 Last accessed 20/04/2019.

*Guardian, The*. n.d., 'Libor-Rigging Emails Lift Lid on City Culture.' Accessed July 30, 2018. https://www.theguardian.com/business/2015/apr/23/libor-rigging-emails-regulators-banks-misconduct Last accessed 20/04/2019.

Guy, Marieke. 2011. "Preserving Your Emails." *JISC Beginner's Guide to Digital Preservation* (blog). March 2,

2011.
https://web.archive.org/web/20110317063716/http://blogs.ukoln.ac.uk:80/jiscbgdp/2011/03/02/preserving-your-emails/ Last accessed 02/05/2019.

Harvard University 2016, Email Archiving in a Curation Lifecycle Context. Harvard University. https://www.youtube.com/watch?v=EFhRP7rjhMM&feature=youtu.be Last accessed 20/04/2019.

Helderman, Rosalind S., and Hamburger, Tom 2016, 'State Dept. Inspector General Report Sharply Criticizes Clinton's Email Practices.' *The Washington Post*, May 25, 2016. https://www.washingtonpost.com/politics/state-dept-inspector-general-report-sharply-criticizes-clintonsemail-practices/2016/05/25/fc6f8ebc-2275-11e6-aa84-42391ba52c91_story.html?noredirect=on&utm_term=.ef3c92f5675a Last accessed 20/04/2019.

Hall, E. 2005. "RFC 4155 - The Application/Mbox Media Type." September 2005. http://tools.ietf.org/html/rfc4155 Last Accessed 02/05/2019.

Howard, Steven. 2011. "Why Preserving Email Is Harder than It Sounds:" Summarized in Chris Prom, *DPC Briefing: Preserving Email: Directions and Perspective*. http://e-records.chrisprom.com/?p=2192 Last Accessed 02/05/2019.

Hyry, Tom, and Onuf, Rachel 1997, 'The Personality of Electronic Records: The Impact of New Information Technology on Personal Papers.' *Archival Issues,* 22: 1. 37–44.

IBM Corporation. 2009. "Demo: Lotus Domino Replication Basics." February 23, 2009. http://www-10.lotus.com/ldd/dominowiki.nsf/dx/domino-replication-basics Last Accessed 02/05/2019

'JSON Meta Application Protocol Specification (JMAP).' n.d. Accessed August 29, 2018. https://jmap.io/index.html Last accessed 20/04/2019.

Klensin, J 2008, 'RFC 5321 – Simple Mail Transfer Protocol.' Internet Engineering Task Force. October 2008. http://tools.ietf.org/html/rfc5321 Last accessed 20/04/2019.

Kim, Sueng Min 2018, 'National Archives Says It Won't Be Able to Produce All Kavanaugh Documents until End of October.' *The Washington Post*, August 2, 2018. https://www.washingtonpost.com/politics/nationalarchives-says-it-wont-be-able-to-produce-all-kavanaugh-documents-until-end-ofoctober/2018/08/02/011a9f1e-966f-11e8-8ffb-5de6d5e49ada_story.html Last accessed 20/04/2019.

Kollewe, Julia 2017, 'Just One in Five Complaints about Potential Mis-Sold PPI Made so Far.' *The Guardian*, March 2, 2017, sec. Money. http://www.theguardian.com/money/2017/mar/02/just-one-in-five-complaintsabout-potential-mis-sold-ppi-made-so-far Last accessed 20/04/2019.

Lappin, James. 2011. "Preserving E-Mail – Records Management Perspectives." *Thinking Records* (blog). August 11, 2011. http://thinkingrecords.co.uk/2011/08/11/preserving-e-mail-records-managementperspectives/ Last Accessed 02/05/2019

Lappin, James 2015, 'How the Cabinet Office's 90 Day Email Deletion Was Reported Back in 2004.' Thinking Records (blog). June 22, 2015. https://thinkingrecords.co.uk/2015/06/22/how-the-cabinet-office-90-dayemail-deletion-was-reported-back-in-2004/ Last accessed 20/04/2019.

Lappin, James. 2018. "How Does Archival Policy towards Email Work?" *Digital Preservation Coalition Briefing Day: Email Preservation "How Hard Can It Be?,"* January 24, 2018. https://www.dpconline.org/docs/miscellaneous/events/2018-events/1769-dpc-email-ii-lappin/file Last Accessed 2/05/2019.

Lazorchak, Butch 2013a, 'The "What" of Email Archiving.' Webpage. The Signal: Digital Preservation | Library of Congress (blog). July 2, 2013. http://blogs.loc.gov/thesignal/2013/07/the-what-of-email-archiving/ Last accessed 20/04/2019.

Lazorchak, Butch 2013b, 'The "How" of Email Archiving: More Launching Points for Applied Research.' Webpage. The Signal: Digital Preservation | Library of Congress (blog). July 18, 2013. https://blogs.loc.gov/thesignal/2013/07/the-how-of-email-archiving-more-launching-points-for-appliedresearch/ Last accessed 20/04/2019.

Mackenzie, Maureen L., 2002. Storage and Retrieval of E-mail in a Business Environment: An Exploratory Study. *Library and Information Science Research*, 24(4), pp 357-372.

Madrigal, Alexis C. 2014, 'Email Is Still the Best Thing on the Internet.' *The Atlantic*. August 14, 2014. https://www.theatlantic.com/technology/archive/2014/08/why-email-will-never-die/375973/ Last accessed 20/04/2019.

Marks, Steve 2015, 'Becoming a Trusted Digital Repository'. *Trends in Archives Practice* 8. Chicago, IL: Society of American Archivists.

Marshall, Catherine C. 2007, 'How People Manage Personal Information Over a Lifetime.' *Personal Information Management*, 57–75. Seattle, WA: University of Washington Press.

Marshall, Tanya 2017, 'Issues in Moving from Policy to Governance: Vermont Email Challenge.' In *CoSAHNPRC Symposium: Government Email in an Age of Risk: Preventing Information Loss*, 6.

Metz, Joachim. n.d., 'Notes Storage Facility (NSF) Database File Format: Analysis of the NSF Database File Format.' notesdesign.com

Meyerson, Jessica, Vowell, Zac, Hagenmaier, Wendy, Leventhal, Aliza, Russey Roke, Elizabeth, Rios, Fernando, and Walsh, Tim 2017, 'The Software Preservation Network (SPN): A Community Effort to Ensure Long Term Access to Digital Cultural Heritage.' *D-Lib Magazine,* 23: 5/6. https://doi.org/10.1045/may2017-meyerson Last accessed 20/04/2019.

Microsoft Corporation n.d., 'Export or Backup Email, Contacts, and Calendar to an Outlook .Pst File.' https://support.office.com/en-us/article/export-or-backup-email-contacts-and-calendar-to-an-outlook-pstfile-14252b52-3075-4e9b-be4e-ff9ef1068f91 Last accessed 20/04/2019.

Ministry of Justice 2018, 'Civil Procedure Rules Homepage, 97th Update.' May 27, 2018. http://www.justice.gov.uk/guidance/courts-and-tribunals/courts/procedure-rules/civil/index.htm Last accessed 20/04/2019.

Murray, Kate, and Engle, Erin 2015, 'We Welcome Our Email Overlords: Highlights from the Archiving Email Symposium.' The Signal: Digital Preservation | Library of Congress (blog). July 9, 2015. https://blogs.loc.gov/thesignal/2015/07/we-welcome-our-email-overlords-highlights-from-the-archivingemail-symposium/?loclr=blogsig Last accessed 20/04/2019.

Myers, J. 1996. "RFC 1939 - Post Office Protocol - Version 3." Internet Engineering Task Force. May 1996. http://tools.ietf.org/html/rfc1939 Last Accessed 02/05/2019.

National Archives, The (UK) 2016, 'Guidance Principles on the Auto-Deletion of Email.' http://www.nationalarchives.gov.uk/documents/information-management/guidance-principles-on-thedeletion-of-email.pdf Last accessed 20/04/2019.

National Archives, The (UK), n.d., 'Managing Emails.' http://www.nationalarchives.gov.uk/informationmanagement/manage-information/policy-process/managing-email/ Last accessed 20/04/2019.

National Archives and Records Administration 2016, 'General Records Schedule 6.1: Email Managed under a Capstone Approach.' *General Records Schedule 26.* National Archives and Records Administration. https://www.archives.gov/files/records-mgmt/grs/grs06-1.pdf Last accessed 20/04/2019.

National Archives and Records Administration 2018, 'National Archives Records Related to Judge Brett M. Kavanaugh.' National Archives and Records Administration. July 31, 2018. https://www.archives.gov/news/topics/kavanaugh-records Last accessed 20/04/2019.

North Carolina Department of Natural and Cultural Resources. n.d. 'TOMES Project | NC DNCR.' https://www.ncdcr.gov/resources/records-management/tomes Last accessed 20/04/2019.

Northeast Document Conservation Center 2007, 'Planning for Digital Preservation: A Self-Assessment Tool.' 2007. https://www.nedcc.org/assets/media/documents/DigitalPreservationSelfAssessmentfinal.pdf Last accessed 20/04/2019.

Owens, Trevor. 2018. *The Theory and Craft of Digital Preservation*. Baltimore: Johns Hopkins University Press.

Palmer, Brian 2010, 'What's a 'Diplomatic Cable'?' *Slate*, November 29, 2010. http://www.slate.com/articles/news_and_politics/explainer/2010/11/whats_a_diplomatic_cable.html Last accessed 20/04/2019.

Paquet, Lucie. 2000. "Appraisal, Acquisition and Control of Personal Electronic Records: From Myth to Reality." *Archives and Manuscripts* 20 (2): 71-91.

Partridge, Craig 2008, 'The Technical Development of Internet Email.' *Annals of the History of Computing*, IEEE 30: 2, 3–29. https://doi.org/10.1109/MAHC.2008.32 Last accessed 20/04/2019.

Pennock, Maureen 2006, 'Curating E-Mails: A Life-Cycle Approach to the Management and Preservation of EMail Messages.' In *DCC Digital Curation Manual*. Digital Curation Centre. http://www.dcc.ac.uk/sites/default/files/documents/resource/curation-manual/chapters/curating-emails/curating-e-mails.pdf Last accessed 20/04/2019.

Postel, Jonathan B., Partridge, C. Klensen, J., Freed, N., Rose, M., Stefferud, E., Croker, D. and Moore, K. 1982, 'STD 10: Simple Mail Transfer Protocol.' Internet Engineering Task Force. 1982. https://tools.ietf.org/html/std10 Last accessed 20/04/2019.

Prom, Christopher J. 2011. 'Preserving Email.' 11–01. DPC Technology Watch. Digital Preservation Coalition. http://dx.doi.org/10.7207/twr11-01.

Purcell, Aaron D. 2016, *Digital Library Programs for Libraries and Archives: Developing, Managing, and Sustaining Unique Digital Collections*. Chicago: Neal-Schuman, an imprint of the American Library Association.

Resnick, ed. 2008. 'RFC 5322 - Internet Message Format.' Internet Engineering Task Force. http://tools.ietf.org/html/rfc5322 Last accessed 20/04/2019.

Rizga, Kristina 2018, 'How to Teach the Civil War in the Deep South.' *The Atlantic*. December 7, 2018. https://www.theatlantic.com/education/archive/2018/12/how-teach-civil-war-deep-south/577588/ Last accessed 20/04/2019.

Schmitz Fuhrig, Lynda. 2011. "Guidelines for Managing Your Work and Personal Email." The Atlantic - Technology. April 28, 2011. http://www.theatlantic.com/technology/archive/2011/04/guidelines-formanaging-your-work-and-personal-email/237961/ Last accessed 02/05/2019.

Schmitz Fuhrig, Lynda, and Nancy Adgent. 2008. "Preserving Historical Correspondence: Email Preservation Progress and Future Directions." Presented at the CERP Symposium, Washington, D. C., November 10. https://web.archive.org/web/20130617090018/http://siarchives.si.edu/cerp/CERP_EMCAP_symp.pdf Last accessed 02/05/2019.

Schneider, Josh 2016, 'Viewing Email through a New Lens: Screening, Managing, and Providing Access to Historical Email Using ePADD.' BloggERS! | Society of American Archivists Electronic Records Section (blog). January 12, 2016. https://saaers.wordpress.com/2016/01/12/viewing-email-through-a-new-lens-screeningmanaging-and-providing-access-to-historical-email-using-ePADD/ Last accessed 20/04/2019.

Schneider, J., Chan, P. and Edwards, G 2017, 'ePADD: Computational Analysis Software Enabling Screening, Browsing, and Access for Email Collections.' *IPres2017 Proceedings*, 4.

Simpson, Joel 2016, 'Email Archiving Systems Interoperability Report.' Harvard Library. https://dash.harvard.edu/bitstream/handle/1/28682572/HL_Email_Archiving_Systems_Interoperability_Report_2016.pdf?sequence=3 Last accessed 20/04/2019.

Scholtes, Johannes. 2006. "A View on Email Management: Balancing Multiple Interests and Realities of the Workplace." *KMWorld*, February 2006.

Smallwood, Robert 2008, *Taming the Email Tiger: Email Management for Compliance, Governance, and Litigation Readiness: A Management Guide*. Original ed. New Orleans LA: Bacchus Business Books. http://books.google.com/books?id=6Q0voyWUf0EC Last accessed 20/04/2019.

Smithsonian Institution Archives 2008, 'The Collaborative Electronic Records Project.' 2008. http://siarchives.si.edu/cerp/ Last accessed 20/04/2019.

Smithsonian Institution Archives 2017, 'DArcMail: Digital Archiving of Email Users Guide.' December 2017. https://siarchives.si.edu/sites/default/files/forum-pdfs/SIA_DArcMail_UsersGuide.pdf Last accessed 20/04/2019.

Stanford University n.d., 'ePADD Project Homepage.' Stanford Libraries. https://library.stanford.edu/projects/ePADD Last accessed 20/04/2019.

Stewart, Emily 2018, 'The Ivanka Trump Email Controversy, Explained.' *Vox*, November 28, 2018. https://www.vox.com/policy-and-politics/2018/11/28/18116326/ivanka-trump-emails-hillary-clinton-gma Last accessed 20/04/2019.

Stripling, Jack 2018, 'Michigan State Chief Said Abuse Victim Would Get 'Kickback' for Stirring Up Survivors, Emails Show.' *The Chronicle of Higher Education*, June 13, 2018. https://www.chronicle.com/article/Michigan-State-Chief-Said/243656 Last accessed 20/04/2019.

Sustainability of Digital Formats 2015, 'Lotus Notes Storage Facility.' Web page. Library of Congress Digital Preservation. December 23, 2015. https://www.loc.gov/preservation/digital/formats/fdd/fdd000433.shtml Last accessed 20/04/2019.

Task Force on Technical Approaches for Email Archives 2018a, 'Guide to Email Standards.' March 7, 2018. http://www.emailarchivestaskforce.org/documents/guide-to-email-standards/ Last accessed 20/04/2019.

Task Force on Technical Approaches for Email Archives 2018b, 'The Future of Email Archives.' New York: Council on Library and Information Resources. https://www.clir.org/pubs/reports/pub175 Last accessed 20/04/2019.

*Telegraph, The* 2013, 'RBS Libor Rigging Emails,' February 6, 2013, sec. Finance. https://www.telegraph.co.uk/finance/newsbysector/banksandfinance/9852632/RBS-Libor-rigging-emails-Itsjust-amazing-how-Libor-fixing-can-make-you-that-much-money-says-trader.-Im-like-a-whores-drawers-addsanother.html Last accessed 20/04/2019.

Tkacova, Monika 2018, 'Accountable and Efficient Disposal of Lotus Notes Data and Databases: A Case Study in Managing Data as UN Records,'. United Nations Archives and Records Management Services.

Tobias, Daniel R. 2017. 'Dan's Mail Format Site.' 2017. https://mailformat.dan.info/ Last accessed 20/04/2019.

United Nations, ed. 2000, *United Nations Correspondence Manual: A Guide to the Drafting, Processing, and Dispatch of Official United Nations Communications*. New York: United Nations.

United Nations Archives and Records Management Services n.d., 'Retention Schedules.' https://archives.un.org/content/retention-schedules Last accessed 20/04/2019.

United Nations Archives and Records Management Services n.d., 'Secretary-General Kurt Waldheim (1972–1981).' https://search.archives.un.org/secretary-general-kurt-waldheim-1972-1981 Last accessed 20/04/2019.

United Nations System Chief Executives Board for Coordination 2016, 'Total Expenditure by Category.' United Nations System Chief Executives Board for Coordination. 2016. https://www.unsystem.org/content/FS-F0004

United States Department of State, Office of the Inspector General 2016, 'Office of the Secretary: Evaluation of Email Records Management and Cybersecurity Requirements.' https://oig.state.gov/system/files/esp-1603.pdf Last accessed 20/04/2019.

United States Securities and Exchange Commission 2003, Final Rule: Retention of Records Relevant to Audits and Reviews. 17 CFR. Vol. Part 210. https://www.sec.gov/rules/final/33-8180.htm Last accessed 20/04/2019.

US Supreme Court 2010, 'Federal Rules of Civil Procedure.' Legal Information Institute, Cornell University. 2010. http://www.law.cornell.edu/rules/frcp/ Last accessed 20/04/2019.

Watson-Tyndall, Camille 2017, 'Working with Stakeholders to Create/Influence Policy: North Carolina TOMES Policy.' In *CoSA-HNPRC Symposium: Government Email in an Age of Risk: Preventing Information Loss*, 6. https://www.archives.gov/files/nhprc/projects/electronic-records/pdf/case-study-1-working-withstakeholders-nc-tyndall-watson.pdf Last accessed 20/04/2019.

Wikipedia. n.d. "Email." 2019. https://en.wikipedia.org/w/index.php?title=Email&oldid=893835866 Last accessed 02/05/2019.

Wikipedia, 2017, 'Cc:Mail', https://en.wikipedia.org/w/index.php?title=Cc:Mail&oldid=762916022 Last accessed 20/04/2019.

Wikipedia, 2018a, 'Comparison of Email Clients', https://en.wikipedia.org/w/index.php?title=Comparison_of_email_clients&oldid=871826762 Last accessed 20/04/2019.

Wikipedia, 2018b, 'IBM Notes', https://en.wikipedia.org/w/index.php?title=IBM_Notes&oldid=852546254, Last accessed 20/04/2019.

Yeh, Jen-Yuan, and Harnly, Aaron 2006 'Email Thread Reassembly Using Similarity Matching.' In *CEAS 2006 – Third Conference on Email and Anti-Spam*. Mountain View, CA. https://web.archive.org/web/20151014032517/http://www.ceas.cc:80/2006/7.pdf Last accessed 02/05/2019.