Preserving Email

Christopher J.Prom

DPC Technology Watch Report 11-01 December 2011

Series editors on behalf of the DPC
Charles Beagrie Ltd.

Principal Investigator for the Series
Neil Beagrie

DigitalPreservationCoalition

DPC Technology Watch Series

Published in association with Charles Beagrie Ltd.

ISSN 2048-7916

DOI http://dx.doi.org/10.7207/twr11-01

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

Foreword

The Digital Preservation Coalition (DPC) is an advocate and catalyst for digital preservation, enabling our members to deliver resilient long-term access to content and services, and helping them derive enduring value from digital collections. We raise awareness of the importance of the preservation of digital material and the attendant strategic, cultural and technological issues. We are a not-for-profit membership organization, and we support members through knowledge exchange, capacity building, assurance, advocacy and partnership. Our 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 that our collected digital memory will be available tomorrow. They provide an advanced introduction to support those charged with safeguarding 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 from experts after consultation among DPC members about shared priorities and challenges, and are thoroughly scrutinized by peers before being released. We ask the authors 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. These reports are a distinctive and lasting contribution to the dissemination of good practice in digital preservation.

This report was written by Chris Prom, Assistant University Archivist at the University of Illinois, Urbana, USA. During 2009–10, as part of his Fulbright Distinguished Scholar Award, Prom directed a research project at the Centre for Archive and Information Studies at the University of Dundee, Scotland, on 'Practical Approaches to Identifying, Preserving, and Providing Access to Electronic Records'. This included a major focus on the preservation of email. The report is published by the DPC in association with Charles Beagrie Ltd. Neil Beagrie, Director of Consultancy at Charles Beagrie Ltd, was commissioned to act as principal investigator for and managing editor of this Series in 2011. He has been further supported by an Editorial Board drawn from DPC members and peer reviewers who comment on texts prior to release: William Kilbride (Chair), Neil Beagrie (Editor), Janet Delve (University of Portsmouth), Sarah Higgins (University of Aberystwyth), Tim Keefe (Trinity College Dublin), Andrew McHugh (University of Glasgow) and Dave Thompson (Wellcome Library).

Acknowledgements

Many people provided invaluable assistance during the drafting of this report. William Kilbride graciously hosted my attendance at a DPC briefing day, 'Preserving Email: Directions and Trends'. I learned much from each of the other presenters – Steve Bailey (JISC), Maureen Pennock (The British Library), Steven Howard (The Special Tribunal for Lebanon), Susan Thomas (Oxford) and Tom Jackson (Loughborough University) – as well as from the audience's searching questions and discussion. I owe Susan Thomas a particular debt, since she supplied information for one of the case studies cited in the text; Crawford Neilson from the Medical Research Council deserves similar thanks.

My colleagues at the University of Illinois, William Maher, Melissa Salrin and Joanne Kaczmarek, all supplied helpful comments on drafts of this report, and graduate student Brandon Pieczko graciously let me see an article under preparation, providing several useful citations. In addition, I had useful and productive discussions about email preservation with Ricc Ferrante (Smithsonian Institution Archives), Ben Goldman (University of Wyoming), Michael Shallcross (University of Michigan), Kelly Eubank (North Carolina State Archives), Tim Gollins (The National Archives, UK), Wendy Gogel (Harvard University), Patricia Galloway (University of Texas), Richard Pearce-Moses (Clayton State University), Donald Post (Imerge Consulting), Claus Jensen (Danish Royal Library) and Hans Boserup (Danish Royal Library).

Neil Beagrie of Charles Beagrie Ltd provided consistent and useful editorial direction, and helped shape ideas presented in the final report; he also developed the report layout and design elements. Finally, I would like to thank my wife Linda and my children Andy, Grace and Molly for their support and patience as I researched and drafted this report over many evenings and weekends. Of course, any inconsistencies, gaps or errors in the final report are solely my fault.

Christopher J Prom
December 2011

Contents

Abstract 1

Executive Summary 1

1. Introduction 2

1.1. The Importance of Preserving Email 3

1.2. Overview of Past Work 5

2. Issues 8

2.1. Technical Challenges Impeding Email Preservation 8

2.2. Technical Factors Facilitating Email Preservation 12

2.3. The Legal Context 12

2.4. Institutional Attitudes and Behaviours 14

2.5. The Personal Nexus: End User Attitudes, Actions and Behaviours 17

2.6. Summary of Issues 19

3. Standards and Technologies 19

3.1. IETF Standards 20

3.2. Proto-Standards for Email Storage 23

3.3. Other Related Standards 24

3.4. Email Preservation Technologies 25

3.5. Email Search, Discovery, Access and Rendering Technologies 28

3.6. Email Preservation Case Studies 29

3.6.1. Bodleian Library: Preservation for Cultural Heritage 30

3.6.2. Medical Research Council: Medium-Term Preservation 32

4. Recommended Actions 33

4.1. For Institutions 34

4.2. For Individuals 36

4.3. For the Digital Preservation Community 37

5. Conclusion 37

6. Glossary and Acronyms 38

7. Further Reading 40

8. References 41

Abstract

Over 40 years after the invention of email, relatively few institutions have developed policies, implementation strategies, procedures, tools and services that support the long-term preservation of records generated via this transformative communication mechanism. However, a close examination of recent literature reveals that significant progress has been achieved in developing the essential elements that can be used to build an effective email preservation programme. By implementing appropriate technical standards, new capture methods and emerging technologies, archivists, curators, records managers and other information professionals working in the cultural heritage sector can take practical steps to preserve email for its legal, administrative or historical value.

Executive Summary

The use of email technologies is increasing and is becoming more embedded in our daily lives. Since email systems generate a constantly evolving yet fragile stream of evidence concerning people's activities, the messages and attachments they transmit can comprise rich documentary research resources. Like the letters of yesteryear, email will be of widespread interest not only to its original senders and recipients, but also to students, scholars and members of the public seeking evidence and information about the past.

In spite of email's potential historical, legal and administrative value, few organizations have developed sustainable programmes that are dedicated to preserving it. Several factors, including perceived technological barriers and legal mandates favouring destruction, have led many organizations pursue policies that amount to little more than benign neglect. As a result, the end users of email systems frequently shoulder the ultimate responsibility for managing and preserving their own email, thus exposing important documentary records to needless and counterproductive risk of loss.

By understanding the technical standards that underlie email systems, by undertaking appropriate preservation strategies and by implementing new technologies, individuals and organizations can lay the foundation for effective email preservation programmes. The following activities will be essential elements of such efforts:

- analysing current email systems vis-à-vis end user needs and behaviours;
- prioritizing long-term access over minimum legal retention periods;
- developing short and easy-to-follow email management guidelines;
- building procedures and tools that make email preservation an integral, transparent and effortless part of email users' communication practices; and
- capturing email in a standards-based, system-neutral format that preserves its significant properties.

As part of these activities, organizational leaders, records managers, archivists, curators and information technology professionals should pursue one or more strategies, as appropriate to local circumstances. They can offer guidance, helping email users engage in

behaviours that support preservation. They can capture individual messages or groups of messages on a case-by-case basis. They can selectively preserve entire email accounts, using client- or web-based migration and capture tools. They can preserve entire email ecosystems by using open source or proprietary email archiving software.

While these approaches will be effective for short- to medium-term preservation, additional work is needed if the community is to effect a long-term preservation environment that supports the capture, storage and access of email. The development of such a preservation environment will be advanced if members of the digital preservation community undertake the following activities:

- articulating the importance of email preservation by demonstrating the value that specific email collections can play in protecting rights, ensuring accountability, documenting events and facilitating research;
- prioritizing the funding of email preservation relative to complementary or competing digital preservation initiatives; and
- initiating research and development efforts to build tools, services, and programs that support the storage, indexing, retrieval, querying and display of email in a self-describing and system-neutral format.

1. Introduction

In 1965, Tom Van Vleck and Noel Morris sent what were perhaps the world's first electronic messages to each other, using the mail function that they developed for the Massachusetts Institute of Technology's Compatible Time Sharing System (CTTS) (Van Vleck 2010). Those who used the mail command in CTTS and its successor system, MULTICS (Multiplexed Information and Computing Service), embraced the technology with fervour (Multicians.org 2011). However, by the late 1960s, these users noted that it was becoming difficult to find information concerning the development of the system because '[t]he [system design] memos seem to have been superseded by email'. The paper documentation for MULTICS – which has survived and been accessioned to the MIT archives – thins out in both quantity and quality after that point (Morris 2011). Even the first email messages exchanged by Van Vleck and Morris have long since gone missing, deemed too trivial to preserve.

Over 40 years later, it is still the case that a relatively small number of institutions have acknowledged that email should be actively preserved for historical purposes. A subset 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.

This report reviews those efforts and offers recommendations for organizational leaders, records managers, IT professionals, librarians, archivists and curators (a group hereafter referred to as information professionals) who are seeking 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. It concludes with a call for leaders in the information technology and cultural heritage communities to develop new standards and tools that facilitate the preservation of email.

This guide does not provide assistance in developing policies or systems to manage email over the short-term, solely for the purpose of complying with laws, regulations or records management imperatives, except to the extent that such efforts may facilitate or impede attempts to preserve email for its historic, cultural or administrative value. Instead, it builds on analysis and advice offered in past work completed by 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.

1.1. The Importance of Preserving Email

A recent survey of Internet usage trends in the United States by the Pew Internet and American Life Project found that email is the most commonly used Internet technology among people of every age group (Zickuhr 2010). In addition, 60 percent of business and IT professionals name it as the single most critical business application (Smallwood 2008, p. 15). Although journalists never tire of proclaiming email's imminent demise (for a typical example from several years ago, discredited by subsequent events, see Lorenz 2007), over 94 percent of all computer-using adults use the technology (Zickuhr 2010, p. 11).

Little evidence substantiates the claim that email use is declining, and several facts refute it. For example, the use of handheld devices to read and send email has risen in an amount exceeding the claimed decline in desktop use (ExactTarget 2010; comScore, Inc. 2011). The use of email on alternative devices illustrates that the technology has become more, rather than less, embedded in people's daily lives. In addition, email usage rises as children outgrow adolescence, assume daily work responsibilities and expand their engagement in civic, social, political, cultural or religious organizations (Ritchel 2010). The percentage of those using email is highest among those between 18 and 33 years old: 96 per cent.[1] As a further sign of its pervasive attraction, personal email use is growing rapidly among the older population, many of whom never used it during their working lives (Zickuhr 2010, pp. 11–15).

Several facts further illustrate email's technological reach. For example, the number of email accounts worldwide continues to grow. Over 3.1 billion email accounts currently exist, and the average business user sends 33 email messages per day (Radicati Group, Inc. 2011a). The existence of resources dedicated to controlling the flood of daily messages,

[1] The Pew study also shows that of people in that age group, 83 percent use social networking sites, 66 percent use instant messaging services and 18 percent use blog authoring software It should be noted that on a time of use basis, 18- to 33-year-olds use email less frequently than their elders.

such as the 'email charter' and the resources provided by Thomas Jackson, testify to the singular importance of this technology in people's daily lives (Anderson 2011; Jackson 2009). Email is the one essential Internet technology that is used by nearly every Internet user.

Recent trends show that email usage is growing and becoming more deeply embedded in our work and personal lives, blurring the line between business and personal communication. One recent study found that 77 percent of email users forward business messages to personal accounts – a practice that not only poses risks to an organization but also complicates email preservation efforts (Smallwood 2008, p. 14). Research regarding people's personal information management practices shows that many people use free online email services as a strategy to keep a record of their activities and to back up their important documents (Marshall 2008a). In addition, a significant minority of email users supplement email messaging with blog posts, text messages or social network updates (for example Twitter and Facebook updates).

Taken as a whole, these trends present an archival paradox. Email and other electronic messages are both ubiquitous and ephemeral, documenting people's professional and personal lives in a chaotic stream of messages and relationships, calling for a new documentary practice (Maher 1992).

Email communications saturate many people's lives and comprise a large portion of Internet traffic, but they are rarely captured in large-scale Internet preservation projects. Of course, a great deal of business and personal activity takes place through blogs, social networking sites, instant messaging programs, video chat services and other electronic messaging systems. The Library of Congress is capturing an archive of all public tweets worldwide. Several agencies provide website harvesting (UK Web Archive 2011; Internet Memory Foundation 2011; Internet Archive 2011; California Digital Library 2011). Blog preservation projects, with significant external funding, are also underway (BlogForever Project 2011). These worthy efforts must continue. But in comparison, email preservation has taken a backseat, with few grant-funded projects and relatively little institutional support.

This is particularly unfortunate because email messages frequently contain evidence of a type and quality that is strikingly different from and complementary to publicly available resources such as web pages, blogs and social media postings (see Beagrie 2005, see Figure 1 and related text). Email generates a constantly evolving yet fragile record of actions taken by organizations or individuals. When an individual sends an email message, the software that executes the transaction leaves an embedded trail of evidence in the email header; this trail can demonstrate how people lived and worked within a network of colleagues, friends and family members. Several articles and projects show the potential for documenting and understanding social networks via email analysis (Sudarsky and Hjelsvold 2002; Viegas, et al., 2004; Perer, et al., 2006; Gorton, et al., 2007; and Stanford University, Mobisocial Laboratory 2011).

Email is particularly valuable because people typically use it to record information that was not intended for wide revelation at the time of sending. Since email communicates private information, it is unlikely that the technologies supporting it will disappear anytime soon. Unsurprisingly, email is frequently the target of hackers, who go to great lengths to procure it. The theft or leaks of email has resulted in two of the more controversial episodes in recent history, the Wikileaks and Climategate scandals (Crook 2010; The Guardian 2011; Leigh 2011). Information contained in email has helped destroy entire corporations, such as the American power trading company Enron and its auditor, accounting giant Arthur Andersen (Hunter 2007). In the long-term, it is likely that the historical value of email messages will only increase, since correspondence is a documentary form that, skilfully interpreted in the light of other evidence, lays bare the sinews of history (Barzun and Graff 1992, pp. 114–17; Vanhoutte and den Branden 2009, pp. 77–78). Email records can be used alongside other types of records to develop complete and nuanced narratives. As Winton Solberg, an eminent historian of American higher education, remarked to the author recently, 'historical research will be absolutely impossible in the future unless your profession finds a way to save email'.

Nevertheless, it is far from obvious to the public at large and even to many members of the cultural heritage community that email constitutes an appropriate object of long-term historical preservation, demanding management for that goal. Information professionals must make the case for preserving it, and in doing so they must address technical, legal and organizational constraints that will colour local attitudes and approaches to preserving email.

1.2. Overview of Past Work

Over the past 20 years, many authors have provided informal advice to those seeking to manage personal email more effectively (Schmitz Fuhrig 2011; Guy 2011; Ashenfelder 2011a). However, relatively few peer-reviewed reports or articles have emerged concerning the topic of email preservation. A few individuals have issued calls for libraries, archives and museums to preserve email correspondence for its cultural value (Hyry and Onuf 1997; Enneking 1998; Marshall 2007; Cox 2008). Others have provided project reports and implementation advice (Paquet 2000; Mackenzie 2002; Schmitz Fuhrig and Adgent 2008; Goethals and Gogel 2010). A systematic review of literature indexing services yielded only four pieces devoted to the technical aspects of email preservation (Li and Somayaji 2005; Milicchio and Gehrke 2007; Srinivasan and Baone 2008; Wagner, et al., 2008a) as well as several publications discussing tangential but important topics such as user behaviour, the history of email, personal management practices or email visualization systems.[2] While these sources provide interesting background reading, the digital curation

[2] See Mackenzie 2002; Billsus and Hilbert n.d.; Beebe 2008; Brogan and Vreugdenburg 2008; Chatelain and Garrie 2007; Cole and Eklund 1999; Gorton et al. 2007; Meyer 2009; Partridge 2008; Perry 1992; Potter 2002; Pukkawanna, et al., 2006; Sudarsky et al. 2002; Viegas et al. 2004; and Whittaker and Sidner 1996.

community lacks a systematic research agenda to establish a theory and practice of email preservation.

Nevertheless, four works provide particularly useful starting points: David Bearman's 1994 article 'Managing Electronic Mail;' Filip Boudrez and Sofia Van den Eynde's 2003 report for the 'DAVID' Project *Archiving Email*; The Digital Preservation Testbed Project's final report, *From Digital Volatility to Digital Permanence: Preserving Email*; 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 works place attention on the entire 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 (Bearman 1994).

Bearman's overall approach should influence current attempts to capture email, although some of his specific recommendations are no longer feasible. In particular, he placed a great deal of faith in the ability of automated tools to filter and capture 'record' emails into the types of electronic records management systems (ERMS) that he believed might serve as an external store for both structured data and for semi-structured records like email. However, the software available at the time he wrote was not fit for the task of realizing his vision.

In the late 90s and early 2000s several institutions attempted to apply Bearman's recommendations concerning email preservation. Reports issued by the City of Antwerp and the Dutch Digital Preservation Testbed Project provide two highly useful resources.

The 'DAVID' report, issued by the City of Antwerp's project team, describes how email systems work, proposes a framework to address legal and ethical concerns over email privacy, reviews policy and technical options, and proposes a set of practical steps to establish an email archive (Boudrez and Van Den Eynde 2002). The authors advance a nuanced discussion of privacy considerations, arguing that plans to preserve email must take cognizance of legal requirements to protect privacy rights, as codified in

telecommunications laws and directives. Boudrez and Van Den Eynde review four options that an institution can use to capture the content, context, and structure of email messages:

- printing to hard copy;
- duplicating all messages transmitted by an email server to a separate archival store;
- providing email users with a shared folder on the server, where they can manually or automatically classify emails based on business function, activity, or subject, or case ID; and
- capturing records outside of the email system, migrating them to suitable preservation format such as PDF or XML and storing them in an ERMS or other preservation system (pp. 34–46).

The authors outline a method by which users can register email as records, classify them into an archival store on the server and transfer them out of the email system for permanent preservation (pp. 47–53). In a companion report, Boudrez provides detailed technical information regarding the operation of the DAVID system as Antwerp put it into practice (Boudrez 2006).

In 2003, the Dutch National Archives Digital Preservation Testbed, working in conjunction with the Dutch Ministry of the Interior and Kingdom Relations, issued a similarly useful report (Digital Preservation Testbed 2003). This includes an extended discussion concerning the properties of a message that must be preserved so that people can judge its authenticity. The report weighs the advantages and disadvantages of three preservation strategies: migrating email to a new version of the software or an open standard, wrapping email in XML formats, and emulating email environments, concluding that using XML to encode email messages is the best long-term strategy for creating a sustainable preservation environment. The final sections of the report recommend a process to convert email to the XML format developed by the Testbed project, listing specific steps that information professionals can take to facilitate preservation. The project also supplied an Outlook plug-in to convert messages to the recommended XML format. Unfortunately, the tool is no longer available at the time of writing, and the project website is only accessible via the Internet Archive's Wayback Machine (Digital Preservation Testbed 2003).

Maureen Pennock's 2006 chapter of the *Digital Curation Manual*, 'Curating E-mails', focuses attention on a range of policy, design and implementation choices that an institution will face in attempting to preserve email. Pennock 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).

These four works offer concise summaries of the legal, technical, organizational and personal barriers to email preservation before presenting general frameworks by which institutions can develop an email preservation strategy. They provide sound advice, but few institutions have found the will or resources to apply the activities that they recommend in a way that is systematic and reproducible. In particular, only a handful of institutions currently preserve email in an XML format, as several of the reports recommend.

This *Technology Watch Report* is an initial effort to suggest practical 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 a trusted digital repository. After reviewing issues, standards and technologies, I will outline practical steps than an organization can take to get started with email preservation, offering specific recommendations for information professionals who are seeking to preserve email for its cultural, legal or administrative value. It concludes by outlining an email preservation research and development agenda that the digital preservation community may wish to pursue, in order to build a more complete set of email preservation services.

2. Issues

Institutions are under increasing pressure to manage the flood of email messages sent and received daily. Each institution must develop its own rationale for email preservation in light of institutional 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 evidence-filled documentary genre.

2.1. Technical Challenges Impeding Email Preservation

Several technical factors make email inherently difficult to preserve. 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.

In addition, individual email messages can and do contain attachments of disparate format and content. Simply capturing and preserving the bits that comprise a message is

challenging enough, but further steps are required if the entirety of the message, including attachments, 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 can entail a degree of complexity far beyond other typical digital preservation activities, such as migrating a homogenous set of documents, images or audio recordings.

In order to understand the basic technical difficulties of preserving email, we must understand how a wide range of hardware and software systems interact to send, receive, and store messages. At base, email is a store and forward technology. The basic technology centres on message transfer agents (MTAs) and user agents (UAs).

The delivery of a message requires interaction between one or more MTAs and one or more UAs. Typically, an email server – such as Microsoft Exchange, Postfix, Sendmail, qmail or Lotus Domino – acts as an MTA. Multiple MTAs move an email message from one computer to another until it reaches its final destination. Once a message has been received by the addressed account, the user accesses the message using a UA, such as a Microsoft Outlook client application, a web-based email application or software on handheld devices. User agents provide a method to view, manage, create and forward messages to one or more designated MTAs. In practice, UAs are client applications directly controlled by a user, and MTAs are email server applications indirectly controlled via a UA.

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, or IETF (Partridge 2008). 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 full headers, bodies, and attachments may be received but not necessarily preserved *in toto*.

From the end user's point of view, message transfer agents do not necessarily interact with user agents 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 user agents 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). It is worth noting that:

- servers may be configured to keep a copy of all received and sent messages automatically (either on the email server or on the user's client computer), but instructions in the client can lead to deletion of those messages;

- servers may enforce a set of routing rules upon receipt, so that certain messages (such as spam messages or messages from a listserv) are automatically discarded or forwarded to other locations;
- client programs that connect to a server usually, but not always, leave the message on the server;
- user agents may or may not replicate a copy of the message to the local device or devices that connect to that server, depending on settings in the server or client software;
- rules established on the server or on the client devices, as well as specific user actions (such as deleting a message or moving it from one folder to another) may remove messages from one or all locations. For example, some server administrators and end users delete or move messages using blunt force (often time-based) rules while others filter messages out using keyword analysis;
- users forward messages at will to other users, where similar actions may take place; and
- backup routines typically restore a user to the last state of the system.

Backups are not a substitute for long-term preservation. Even if incremental backups of email accounts are being completed, many messages may never enter the system or may be quickly purged from backup as old backups are overwritten. Additionally, 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 each location, the content of the email message may be stored in a format that differs significantly from that stored elsewhere. In spite of this replication, email is extremely susceptible to loss through deliberate action, user error, or malfeasance (Fallows 2011).While it may seem like simple matter to preserve a single messages or group of messages, the entire email ecosystem for even one user can become extremely complex. Therefore, any programmatic attempt to preserve email must begin not only with an understanding of the specific technologies used in an email ecosystem, but a detailed knowledge of how server administrators and end users have configured software and hardware.

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

1. **Preserving the significant properties of email.** The InSpect Project (Investigating the Significant Properties of Electronic Content Over Time) identified 14 message header properties and 50 message body properties that in ideal circumstances should be retained in order to preserve the authenticity and integrity of email (Knight 2009). While it is likely that many of the email migration tools discussed in section 3.4 of this report will preserve these properties, additional testing needs to be done with common email migration tools. Such testing would help an institution

decide if a particular tool meets local needs, given the way that different systems record header values and structure the bodies of messages.

2. **Context/Threads.** When users respond to a particular email message, they may not include all the relevant information from the message to which they are responding. Sender or recipient information may be a mere stub. In addition, different email servers thread related messages using variant protocols, some using non-standard 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).
3. **Attachments.** It is relatively easy to preserve the bytes that make up an attachment, either as they were encoded upon sending or in its 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 types, 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.
4. **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.
5. **System Documentation.** Many of the existing email systems are implemented with poor or at least difficult-to-use documentation. Most of the open source email servers and clients contain online documentation for end users, but technical matters, such as header syntax and storage formats/structure, tend to be under-documented. In addition, server administrators make many choices when installing or managing an email server, and these decisions can drastically affect how or whether messages are preserved. Unfortunately, these choices are rarely documented.

While these issues cannot be treated in detail, each of them should be discussed in light of local circumstances as an email preservation programme is developed. The tools discussed in section 3.4 provide features which can partially address these points.

2.2. Technical Factors Facilitating Email Preservation

In spite of the technical challenges noted above, efforts to capture and preserve email are facilitated by a simple fact: the message exists in a standardized format at the point of transmission. The ways in which MTAs and UAs operate means that the complete content of individual messages, including headers, bodies and attachments, can easily be captured at that point in time.

Email capture is facilitated by another fact: essential header values are recorded in a structured, well-documented fashion. Each time the user hits the send key, his or her activity is documented as metadata in an email header, showing exactly what was written, to whom, and at what time. Although is far from impossible to forge header metadata or to modify header metadata after the fact (if the user has a deep familiarity with email format and some advanced computer skills) much email is stored on central servers where it cannot be modified by end users. Since the general trend is toward increased use of server-based storage, the header is difficult if not impossible for the casual user to modify, and it typically resides in a format that allows for ready capture or migration, along with the unstructured or semi-structured information contained in the body and attachments.

The increased use of server-based storage and the adoption of cloud-based email services such as Gmail and Hotmail therefore serve as effective weapons in the battle to capture email. Messages for one account are likely to be stored in one place and replicated outward, provided that IT managers have not configured the system with hard storage limits or rigid auto deletion policies. While messages on the server are subject to deletion through deliberate action, error or malfeasance, the existence of centralized sending, receiving and storing servers offers information professionals a potential capture location. This additionally leads to the widespread replication of messages on local devices, from which they may also be captured – provided they have been effectively managed and saved by the end user on those devices.

2.3. The Legal Context

The fact that email headers, bodies and attachments contain evidence makes messages very interesting to government officials, lawyers and anyone else trying to ensure accountability or uncover misdeeds. Therefore, the legal context in which email messages subsist should be of significant concern to anyone seeking to preserve it. Laws and regulations are particularly important to cultural heritage institutions, since they shape institutional practice, leading to the destruction or preservation of messages that may have historical value.

Maureen Pennock has provided a detailed overview of the UK legal environment (Pennock 2006, pp. 11–14). The purpose here is not to repeat Pennock's analysis, but to explain how recent legal developments affect the way in which different types of institutions manage email on a daily basis. With the exception of the Scotland, the UK legal environment has not substantially changed since 2006.

Unfortunately, legal requirements to retain email of historical value have not always been implemented effectively, as the now notorious example of email mismanagement and subsequent lawsuits concerning White House email demonstrated as early as the 1980s (Bearman 1993; Gewirtz 2007; Cox 2008, pp. 215–16). The embarrassing leak of records from the University of East Anglia, England (Crook 2010; Levsen 2009), as well as ongoing legal actions over sloppy records management practices for email in the private sector, testify to the fact that email management problems continue to plague many institutions (Pinguelo and Gonnello 2010).

If email often contains evidence and information that holds 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. In this respect, three areas of law and regulation have been particularly influential:

1. **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. Email therefore must be managed in accordance with the principles of the prevailing law and best professional practice (Pennock 2006, p. 10; Baron 2010). For example, since the passage of the 2011 Public Records Act, all public records in Scotland must by managed under a records management plan. It is important to note that, at least in the United Kingdom, universities are not subject to public records laws, but may use records schedules and plans as good professional practice and to ensure compliance with other laws, such as FOI. FOI laws, which affect many public institutions, might be seen as encouraging email preservation and accessibility, at least for the short term. However, some experts recommend that institutions should not create an email trail in the first place (Smallwood 2008, pp. 30–31, 97–103), so as to avoid the necessity of complying with FOI requirements.
2. **Privacy, data protection, financial oversight laws**. These laws outline records retention compliance periods or set strict rules regarding data use. Once these periods have been met, any further retention presents a discovery risk to the organization (Scholtes 2006a). As a result many organizations, particularly in the private community, are advised to keep messages only as long as legally necessary: an active inducement to deletion as a risk management technique.
3. **Rules of Civil Discovery.** Changes to the US Federal Rules of Civil Procedure (FRCP) took effect on 1 December 2006 and were codified in 2010 (Yoshinaka 2007; Swartz 2006; Juhnke 2003; US Supreme Court 2010). They defined the concept of Electronically Stored Information (ESI) and established rules under which ESI must be provided. 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 (Yoshinaka 2007; Swartz 2006; Juhnke 2003). The equivalent discovery rules for the UK are far less

prescriptive (Foggo, et al., 2007; Ministry of Justice 2011), but the US legal regime affects any company doing business in the United States, including UK companies. In addition, the US rules changes have encouraged the development of a market for email archiving software, as discussed in more detail in section 2.4.

Taken as a whole, these laws and regulations have encouraged at best a passive attitude toward email preservation. At worst, they provide an inducement toward active destruction after the prescribed compliance period has passed.

On the other hand, government agencies and other bodies with public funding (such as many universities) have an interest in the long-term reuse of records. Policies mandating the retention of records can be formalized into a records schedule, which may include review by an archivist and some attention to historical 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.

Assuming that email can be captured and saved, there is an additional legal area that will need attention as an email preservation programme is developed – copyright and intellectual property laws. The copyright status of a message will affect what can be done with it over the long-term. An institution may make a plausible claim to own copyright only for records written by a member of that organization in his or her official capacity or for records donated by a private party where copyright is conveyed in the gift agreement. Of course, every set of email is likely to contain records sent by third parties, over which the organization or donor does not own the copyright. Particular sets of email may contain private data, such as medical or health information, which the sender did not anticipate making public. In case of email donated by a private party, the donor may request or demand a restriction. For these reasons, institutions must develop email access and redaction policies or other restrictions over materials. Such policies must be clearly documented, transparent, and impartial, so that they can be implemented in procedures, services and hardware or software tools. Institutions and individuals wishing to preserve email will need to take active steps to manage the intellectual property rights of email authors. Some guidance regarding this topic can be found in the forthcoming *DPC Technology Watch Report: Intellectual Property Rights for Preservation*, by Andrew Charlesworth, as well as in his legal and ethical issues guidance for the Digital Lives project (Charlesworth 2009).

2.4. Institutional Attitudes and Behaviours

After gaining an understanding of the technical and legal 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 preserving. As Maureen Pennock noted:

> To date, most institutional activities have focused on the management of e-mail messages and have yet to progress beyond this, despite the emergence of a number of XML-oriented solutions. Whilst the technical challenges of implementing an e-mail curation strategy are by no means wholly resolved, the organizational and cultural challenges remain a significant barrier. (Pennock 2006, p. 37)

In particular, perceived storage costs can be a barrier to effective email preservation, at least in some organizations. IT managers may face pressure to limit the amount of email that users can store in their personal quota.

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 will identify and keep those records requiring long-term management, while allowing the deletion of trivial or 'non-record' items. 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 a headache for current IT staff charged with storing the ever-growing bulk of messages, whilst leaving the archivist with a future problem of even greater proportions: an unsorted mass of messages that can only be disentangled via imprecise technical algorithms or expensive hand work.

One popular guide to email management, for example, proceeds from the assumption that institutions should strive to:

- produce as much 'record-free email' as possible by using a messaging service;
- delete all messages in individual accounts as soon as possible, using a blunt chronological delimiter; and
- manage any 'record' email within an integrated Electronic Document and Records Management System (EDRMS). The EDRMS would strictly enforce classification rules and retention periods, deleting records as soon as possible after meeting compliance requirements (Smallwood 2008, pp. 30-31, 97-103).

However, Steven Howard and James Lappin have noted that most attempts to capture and preserve email in EDRMS systems have not proven themselves to be effective. In any case, the marketplace for EDRMS software seems to be contracting, replaced by enterprise content management software typically used to help institutions comply with legal and regulatory requirements (Howard 2011; Lappin 2011).

In practical terms, many institutions simply set quotas or chronological limits, leaving email management to individual employees (often within the context of a difficult-to-enforce email policy). Others outsource email to a cloud service, such as Google's mail service. At least in the United States, such decisions can expose email to potential legal disclosure,

without the records creator being served a subpoena or even informed of the emails' release (Gruenspecht 2011). Other organizations either let email boxes grow to very large sizes, trusting the user to manage his or her records, or they set very low quotas and expect users to save messages to local folders. As a variation on this theme, many records managers and IT professionals set the email server to delete all messages of a defined age, cap email storage space, or remove accounts as soon as an employee leaves. The presumption behind such actions is that email should be deleted from institutional servers as soon as possible, since it is a risk and takes up a lot of room.

While some of the approaches listed above stop short of doing nothing, they place all responsibility on end users, requiring them to take a specific action if they want to save a message. Users can forward email to a personal account, save it in a folder on the local computer or a shared drive, or upload it to an EDRMS. Whatever implications these actions might have for the institution's risk management, they pay little cognizance to historical or long-term administrative value of such records. The bottom line is that they leave decisions about significance to the discretion of the end user (Cox 2008, p. 233), forcing the archivist or records manager into a position of working with individuals on an ad hoc basis, if there is to be any hope of preserving email.

One former academic noted privately to the author that:

> [M]any universities have simple rules, mechanically enforced about account deletion after people have left. I have fought unsuccessfully against this, particularly when I was in [name of city]. The net effect, as far as I am concerned, is policy-led automated deletion of institutional records. [The] response was, basically, that they didn't care about emails as records (apart from, say, the Principal's emails); they only cared about committee decisions (and to be fair, they put a lot of innovative effort into those). Sure enough, a month or so after I left [city], my email account was removed. Likewise, 6 months after I left [another university], that email account was removed. So all those records are effectively gone.
>
> Except they aren't. Somewhat against various policies that I've signed, and probably against the Data Protection Act and others, I've kept a copy of all my key emails […]. I omitted to copy [some] emails, and when I wrote to my friend and erstwhile colleague . . .asking if he could find them from a backup tape, it was literally the week after they had been deleted in a tidy-up. This 'keep everything yourself' policy used to be completely impractical, but since around 1995 has been perfectly fine, as each successive computer generation comes with a hard disk that allows all the files from the previous computer to fit in a fairly small corner.
>
> That does leave me with the problem of ensuring I don't lose my copy. I'm doing my best with backup, including posting the occasional backup to an off-site relative. My next responsibility will be what to do with this lot when I die. It's possible, even likely, that no one is interested. I'm not sure how I'll find out!

The emergence of an 'email archiving' software market, valued at 2.5 billion USD in revenue in 2011, suggests some businesses are now beginning to manage email outside the existing email server infrastructure, as companies start to worry about legal risk (Radicati Group, Inc. 2011b). Email archiving software captures every sent or received email to an external store, where it cannot be deleted by the end users. It also allows IT staff to set retention periods based on defined criteria, such as sender, recipient, subject, keyword or date. Generally, institutions delete email when legal retention periods have passed. In light of the possibility of legal sanctions and the large cost of EDRMS systems, some companies use this 'save it all' approach, applying legal holds over all records when a lawsuit is filed or suspected. Other vendors apply automatic classification at point of capture, based on keyword selection, and their systems allow for different retention periods at very granular levels (by classification, date, sender, recipient, subject and so on). Email archiving applications offer institutions a viable method of preserving messages over the medium term, since they provide a capture and hold mechanism that places messages outside the business system supporting daily email use.

In summary, organizations are under legal pressure to delete email as soon as the legal need to retain has passed. Traditional records management approaches, which focus on identifying and saving 'record' email, have proven very difficult to implement, for both technical and organizational reasons. A number of organizations are beginning to use email archiving software, but this is most often configured to delete all the captured messages after relatively short retention periods. Of course, personal email, such as that kept in Gmail accounts, would not be covered by such systems. Therefore, much of the responsibility for long-term email preservation, for both professional and personal accounts, currently lies at the fingertips of end users.

Richard Cox and many other authors argue that archivists, records managers and IT administrators should work together to develop policies and procedures to manage email for its long-term value (Cox 2008). For the reasons discussed above, such policies will be ineffective unless they are also developed in close collaboration with the email system's end users.

2.5. The Personal Nexus: End User Attitudes, Actions and Behaviours

Any programme to preserve email for its long-term value must proceed from a solid understanding of end user preferences and behaviours. As Steve Bailey has noted, the behaviours that people demonstrate while using email programs have not been studied systematically. The need to meet external mandates has driven records management theory and practice much more directly than meeting end users' information needs (Bailey 2011a; Bailey 2011b). Nevertheless, research conducted over the past several years concerning personal information management sheds quite a bit of light on how users manage their email within existing policy, IT and personal environments.

The rise in computing power and cheap storage has led many people to compile large personal digital libraries, often including email (Beagrie 2005). Some people care deeply about preserving their email and even go to extreme lengths to do so (Cavender 2010). Others simply keep a copy of all emails that they send or receive. From a preservation point of view, the best way to do this may seem to be keeping the messages on the central server, retrieving items from that server as needed. However, this method also poses the risk that once a message is deleted, it may be all be impossible to retrieve it. Recent research shows that some people use email programs to organize their digital records and make them searchable; this lays out a feasible strategy by which people can preserve a digital record of their lives, via their email accounts (see Whittaker, Bellotti and Gwizdka 2006).

More typically, users follow a policy of benign semi-neglect, assuming that the haphazard backup strategies that they employ will enable them to preserve their important documents, email included. According to Cathy Marshall, people use six flawed strategies to implement this 'archiving instinct'; in one of the more common strategies, people email documents to themselves to create an ad hoc archive, in an attempt to 'communicat[e] with a future version of oneself' (Marshall 2008a). Unfortunately, email stores easily degrade or get lost over a lifetime (Marshall 2007). Most free and paid personal email services offer no long-term service guarantee. They rarely even promise to provide the user with a copy of his or her email if the service is terminated. As a result of these trends, a variety of authors provide advice on 'digital estate' planning (Ashenfelder 2011b; Carroll and Romano 2011).

Archival approaches to email should support the preferences that people have expressed for personal digital archive applications and the fact that people often prefer to manage their resources by scattering copies among multiple locations (Marshall 2008b). Email capture and preservation applications should provide passive savers with easy to use methods to record login/passwords in a secure fashion, to bequeath those passwords to others and to aggregate copies of distributed email in one trusted location. To support those who like to actively control their files, the application should also allow individuals to exercise direct control over single messages or groups of messages; to apply value assessments based on source, activity level or recommended disposition; and to catalogue email stores.

In respect to user preferences, any approach to preserving email must acknowledge that people use email technologies in very different ways. For example, many email users, both in their professional and personal lives, avail themselves of a constantly evolving array of services and tools that blur the line between email, other messaging services (such as instant messaging and voice mail), blogs, social networks and business tools (such as customer relations software). For example, many blogging platforms, commenting systems and social media services now allow users to post to other services directly from an email account. Some blog plug-ins allows users to email a backup copy of their data directly to an inbox (Moidu 2009; Gregory 2010; Croxall 2010; ExactByte, LLC 2011). While capturing

email would not preserve a complete record of a person's digital activities, it may prove to be the central activity on which we should be focusing most attention, since many of other services (such as blogs, commenting systems, bulletin boards and social media sites) leave a trail of footprints in email accounts. Ideally, the cultural heritage community would focus efforts on developing an integrated personal archiving toolset, which would help individuals save records of various types in a trusted, centralized, location.

In any case, these trends demonstrate that people's personal information management practices deeply affect how email is managed on a daily basis and how it might be captured and preserved. As the Paradigm project demonstrated, the principles used to preserve an individual's 'papers' in a manuscript collection hold much relevance in the digital realm (Universities of Oxford and Manchester 2008). The case study cited in section 3.6.1 of this report provides some evidence of how the personal touch in capturing and preserving email can be highly effective.

2.6. Summary of Issues

Email systems are simply communication utilities that can be used to send any kind of content, and email is the single 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.

However challenging these facts may seem, they do not mean that email preservation is impossible or even difficult. The emergence of email capture and archiving programs provides us with powerful tools – provided we know to take advantage of them.

3. Standards and Technologies

By understanding current standards, tools and services, individuals and institutions can lay the foundation for a project or programme to capture, store and manage email for long-term preservation. Before beginning a preservation programme, one must understand two important technical details regarding the sending, receipt and storage of electronic mail messages: 1) email transmission is completely standardized around an open standard; and

2) processes surrounding receipt and storage are standardized only to a very limited extent. 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.

3.1. 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. Email standards are defined in a series of 'request for comment' documents issued by the IETF (Tobias 2011). The latest such document, RFC 5321, prescribes the manner in which mail is transported across the Internet: the Simple Mail Transfer Protocol, or SMTP (Klensin 2008). Under the requirements of SMTP, MTAs relay a message from one networked computer to another until it reaches its destination. The headers of the message record the address of the sending and receiving computers, information provided by the user or user agent, and tracking information recorded by MTAs that touch the message while it is in transit. Influenced by obscure engineering decisions, SMTP evolved slowly and in fact the standard remains in draft form today.

SMTP mandates that when MTAs are moving messages, they must subsist in a standardized, ASCII-based format. In RFC 5322 and its predecessors, IETF specified the format for the message bitstream: the Internet Message Format or IMF (Resnick 2008). IMF and related standards do several things:

- require that a message include headers and bodies;
- list the names of, define the expected content for, and specify the number of times 21 pre-defined headers may be included in one message;
- provide for the inclusion of additional headers, at the discretion of particular MTAs
- specify a general syntax for all headers;
- mandate the use of two headers ('orig-date' and 'from');
- recommend the inclusion of three headers ('message-id,' 'references' and 'in-reply-to'); and
- impose requirements on some headers (for example, 'message-id' must be globally unique, if included).

These standards mandate a particular syntax for the message headers and body, in cases where the message consists of a single part and contains only ASCII data. It is important to note that the header typically includes much of the information that can help users judge the authenticity of messages, including timestamps, recipient lists and transmission paths. In addition, headers may contain user classification data, such as folder titles. This essential metadata must be preserved, or the record's integrity may be severely undermined.

A series of companion RFC documents, collectively known as Multipurpose Internet Extensions (MIME) define how multipart messages can be formatted, as well as how non-ASCII-based content (for example, non-Latin character sets) and binary files (such as images, documents or other files) can be encoded within a multipart message (Brodkin 2011). The existence of content in MIME format is noted in the message header by the inclusion of the header 'MIME-Version: 1.0'. The message header also includes a content-type header, specifying the boundaries for the section or sections of the message that contain the MIME content. Each section defined as containing MIME content includes one or more sections of MIME-encoded data, as well as header information describing each part. MIME headers include:

- 'content-ID,' providing a globally unique identifier for the MIME content. Not every email server includes content IDs and formatting is at the discretion of the server;
- 'content-type,' describing the generic data type that is included (for example, text, image, and so on), and specifying the boundaries that mark the beginning and end of the MIME-encoded content;
- 'content-disposition,' instructing the receiving application as to whether the file should be displayed inline or as an attachment; and
- 'content-transfer-encoding,' specifying whether a binary to text encoding protocol has been used, and if so, which one.

Again, most if not all of the information in the MIME header must be preserved if the particular attachment is to be preserved and rendered. For example, the content-type header information could be used to identify particular file types needing special viewing software or requiring future preservation actions, such as migration. Similarly, if the encoding information for an attachment is lost, an email client will lack sufficient technical metadata to convert the MIME content to its native binary file.

SMTP, IMF and MIME regulate how messages are sent from one MTA to another. When an MTA (which is part of an email server) sends messages using this suite of protocols, the receiving server can easily process them.

A companion set of standards, the Internet Message Access Protocol (IMAP) and the Post Office Protocol (POP3), specifies how user agents may connect with email servers to allow an individual to view, create, transfer, manage and delete messages (Crispin 2003; Myers 1996). As noted earlier (see section 2.1), the ways in which IMAP and POP3 operate determine where copies of messages will be stored. Generally speaking, messages accessed via an IMAP connection will be easier to preserve, because copies will most likely be retained on the server and may also be replicated to local devices. On the other hand, messages accessed via a POP3 connection will likely be deleted from the server and kept only on one client device. 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. In addition, particular server/client combinations may use proprietary protocols. For example, Microsoft Exchange Servers interact with Outlook via the Message Application

Programming Interface or MAPI (Microsoft Corporation 2011); similar protocols exist for the IBM/Lotus Domino and Novell First Class servers. However, those servers also support the IETF standards. Finally, it is very important to note that use of POP3 has been declining, since once the message is deleted from the server, it cannot be read on another computer or device, and many users wish to access messages using multiple devices.

Today, nearly every email server and client supports these IETF standards, including proprietary MTAs such as Microsoft Exchange, Lotus Domino, and Novell First Class, as well as open-source servers such as Postfix, Sendmail and qmail. Although we do not need to delve into additional detail regarding these standards, three facts are worth noting because they affect the long-term survivability of the content, context, and structure of messages.

First, 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 the 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 email-threads, 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.

Second, the IMF serves as the basis for two of the most common storage and exchange formats: MBOX and EML (Hall 2005; Wikipedia n.d. 'Email'). In reality, neither of these methods constitutes a storage format per se, since each server or client implements them somewhat differently. 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, with MIME content stored directly in the file. Files can and do grow to astronomical sizes, and even slight file corruption may affect the ability of certain clients to access individual messages or even the entire folder. MBOX files also include the attachments in their MIME format, meaning that action will likely need to be taken to migrate them, if they are to remain accessible in the future. EML files typically store each message as a single file, and attachments may either be included as MIME content in the message 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. Most modern email clients and servers can import and export one or both of the formats, and programs such as Aid4Mail and Emailchemy can readily convert between the two formats and numerous proprietary formats. These programs also save attachments as external files, with a pointer in the original message.

In the case of many 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 in section 3.4 can then convert these files to MBOX or EML. Similarly, an institution might come across email that originated in an obsolete system. Extraordinary efforts may be necessary to migrate such email. In this case, tools such as Aid4Mail, Emailchemy or Xena can convert many file types to MBOX or EML formats. As the Oxford case study cited in section 3.6.1 shows, 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, such as the FTK toolkit, to access the files.[3] In general, if an institution can get email into one of the MBOX or EML formats, it has taken a very big step on the road toward preserving email.

Third, removing messages from their native systems may negatively affect the ability of systems to search, discover, retrieve or render them. This is not to say that messages will become completely inaccessible; in many cases the benefits of format neutrality can be balanced against the risk of keeping email in its native format. The XML conversion tools discussed below can be very useful in achieving format neutrality. However, the author is aware of no general-purpose tools that are intended to facilitate the access, display, searching, or visualization of messages that have been migrated to XML. Until such tools have been developed – if they ever are – institutions will be forced to provide access to migrated messages using an email client of their choice or the user's choice, recognizing that specific tools support different functionality.

3.2. Proto-Standards for Email Storage

As Maureen Pennock and several others have noted, XML offers an attractive option for storing and potentially managing and providing access to email messages (Green, et al., 2002; Potter 2002; Scholtes 2006b; Baron 2010; Goethals and Gogel 2010). Several institutions have developed XML formats to store messages or account information (Boudrez 2006; Carden 2011; Klyne 2003; Library of Congress 2010; Minor 2008; Digital Preservation Testbed 2003). In theory, using an XML format should facilitate the long-term preservation of message content by allowing organizations to write email messages into a self-describing file or files.

The Email Account Schema, an XML format developed by David Minor and Steve Burbeck, currently provides an excellent initial implementation of such a storage format (Smithsonian Institution Archives 2008; North Carolina Office of Archives and History 2009). Jointly sponsored by the Collaborative Electronic Records Project (CERP) and the Electronic Mail Collection and Preservation (EMCAP) project, it is used by the Smithsonian Institution Archives, several US state archives and the Rockefeller Archives Center. Harvard University is also considering its use (Goethals and Gogel. 2010). In general, the standard is very well designed, supporting the encoding of extended headers using paired <name> and

[3] For more information regarding the use of forensics software in preserving digital content, see the forthcoming *DPC Technology Watch Report: Digital Forensics and Preservation*, by Jeremy Leighton John.

<value> elements, preserving the complete content of an account and allowing multiple options for handling MIME content. Attachments can either be encoded in the xml file itself or written in their original binary formats to externally referenced locations. The latter feature is particularly useful because the preservation of the attachments may require additional effort, including monitoring for format obsolescence and the development of future migration actions. A tool to convert MBOX files to the XML account schema format is described in section 3.4.

Institutions are beginning to implement the Email Account Schema, but few tools exist to query, display and render messages that are stored in the format. If the digital preservation community were to develop tools that support the Email Account Schema or a different XML standard for email, that XML format would be a likely candidate for adoption as an International Council on Archives or even an ISO standard. It could be adopted via a process similar to that which led to the development of the PDF/A. Standards development would also be greatly facilitated by the development of applications that allow users to search, discover, view and visualize messages that are stored in the XML format.

Until such applications are developed, it may seem that there is relatively little immediate benefit to be gained by migrating email into an XML-based format. Therefore, institutions that decide to keep email in an XML format should also keep a copy of messages in one of the IETF formats, preferable EML, since it allows attachments to be written as separate files. Repositories should closely track the location and formats of attachments, so that they can search, retrieve and display messages using tools such as those noted in section 3.5.

3.3. Other Related Standards

Once email has been converted to a storage format, institutions should store the files securely, implementing reasonable access controls in line with requirements from institutional policies or donor agreements. In particular, institutions should strive to maintain the authenticity of the files by recording descriptive and preservation metadata for the files and by placing them into a trustworthy digital repository. The DRAMBORA toolkit and the Trustworthy Repositories Audit Certification guidelines offer guidance and help (DRAMBORA Project n.d.; Dale and Ambacher 2007). In addition, repositories should develop a long-term preservation strategy, in line with the requirements of the Open Archival Information System Reference Model (Consultative Committee for Space Data Systems 2002) or the functional requirements specified by the InterPARES project (InterPARES Project n.d.).

These standards imply that repositories should design software and hardware systems and implement them with capable people and adequate funding, in order to fulfil the following email preservation functions:

- identifying file types for attachments;

- recording and auditing file fixity information (checksums);
- generating archival information packets that include the email messages, attachments, associated descriptive and structural metadata, fixity information, and representation information;[4]
- recording the existence and location of non-email records generated by the same records creator;
- storing archival information packets, preferably replicated to two or more physical locations;
- securing files against accidental or purposeful deletion, alteration or tampering;
- monitoring the file format of email messages and related attachments, migrating them to new formats where appropriate; and
- providing dissemination copies to users, in compliance with access policies.

3.4. Email Preservation Technologies

Many resources describe and evaluate software that can support general digital preservation work (see Library of Congress 2011; National Library of Australia 2011; Open Planets Foundation 2011; CJ Prom 2010). In general, people wishing to preserve email should become skilled in using multiple email clients (such as Thunderbird, Exchange, or Apple Mail) to import and export messages, building a skill set for working with old email clients and migration software. Once such skills have been developed, several types of tools can be used to support email preservation work, in three functional areas:

1. **Tools supporting capture at time of transmission or receipt.** 'Email archiving' applications include both open-source and proprietary software and/or hardware. These tools capture the text stream at point of transmission, saving it to an external store. More sophisticated packages include the ability to filter messages on capture, perhaps using a 'miltering' (that is, mail filtering) technology, to browse and search the store, and to apply retention periods and audit rules to messages. Examples include Symantec Enterprise Vault, Iron Mountain Nearpoint, Smarsh Email Archiving Suite, and Mail Archiva (a dual-licensed program, discussed in more detail below). Generally speaking, such software requires professional systems support and maintenance. Several resources provide reviews or evaluations, which would provide essential help when making an implementation decision (Hill 2011; Harbaugh 2010; Harbaugh 2011). 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.

[4] For common attachment types, representation information may be provided by file extension associations established on the server or client machine. Ideally such information would be recorded in the archival information packet.

2. **Tools that support the migration of email from one storage format to another.** Email migration software, such as Aid4Mail, Emailchemy, MessageSave and the CERP Email Parser can read email in one or more defined formats and save it in one or more defined target formats. The most sophisticated 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 EMCAP software provided through the website of the North Carolina State Archives, is an open-source, Windows and SQL Server-based application. It can read emails from an active Exchange account or .pst files, transferring them to a designated archival server (North Carolina Office of Archives and History 2009).
3. **Tools to manage email within an EDRMS or Content Management System.** These systems, which typically provide an organization with a means to comply with records management requirements, may include a method to declare, register or classify mail outside of the email system. Prominent examples that include an email module are HP Trim and the dual-licensed Alfresco software, discussed in more detail below. 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. However, such systems can be useful in highly centralized or litigation-sensitive industries and agencies.

To date, relatively few applications in these three functional areas have been tested to see if they preserve the significant properties of email, as identified by the InSPECT testing report; testing other applications might constitute part of a research agenda regarding email preservation. Nevertheless, the author believes that the following applications would be particularly useful additions to the email preservationist's toolkit:

Adobe Acrobat Pro. General office applications, such as Adobe Acrobat, may play a limited role in email preservation projects. When Acrobat Professional has been installed on a local workstation that also has Microsoft Outlook, a menu item will be added to Outlook, allowing users to save individual messages or groups of messages to a PDF file or PDF portfolio. However, messages saved in these formats will experience a loss of fidelity. In particular, portions of the header will be excluded, and attachments will be encoded (in an unspecified format) inside the PDF file. Such a roundabout method of preservation poses risks. Nevertheless, conversion to PDF may have some limited use, if an organization is unable to secure a copy in a better format for personal, political or administrative reasons, and in particular if the messages do not contain attachments or dynamic content.

Aid4Mail. A desktop application, Aid4Mail can convert many mail formats to a wide range of open and proprietary formats, with a high degree of fidelity, based on informal testing completed by the author of this report. Aid4Mail can also connect to IMAP and MAPI servers (such as Gmail and Microsoft Exchange) to harvest email directly. It includes a filtering system to exclude or include messages meeting stated criteria, a scripting language to allow for custom export formats, and the ability to save emails directly to

PDF/A format. It is a very stable program, quickly converting even very large accounts, and it is available with an 'Archivist' licence tailored for cultural heritage institutions seeking to preserve accounts for historical purposes. (See http://www.aid4mail.com for more information.)

Alfresco. A web-based, dual-licensed content management system that complies with MoReq and DOD 5015.2 requirements, Alfresco provides a method for end users to transfer stored messages into an EDRMS system, declare emails as records, and to manage them under retention policies. Once a system administrator has connected Alfresco to an IMAP server, users can transfer emails into the system where they will reside with other records, such as desktop application files. Using a web-based dashboard, users and administrators can classify messages and apply retention rules to them. See http://www.alfresco.com for more information.

CERP Email Parser. A web application that runs in an open-source virtual machine (Smalltalk Squeak), the CERP Email Parser will transform single or multiple MBOX files into one XML file holding the contents of an entire email account, complying with the requirements of the XML Account Schema Format. Under the leadership of the Smithsonian Institution Archives and the North Carolina State Archives, the CERP project partners collaboratively developed the parser, which can be downloaded from the project website (http://siarchives.si.edu/cerp/). The application allows several options for attachment handling and encoding, supporting both embedded and externally referenced attachments. According to project leader Riccardo Ferrante of the Smithsonian Institution Archives and Kelly Eubank of the North Carolina State Archives (whom the author interviewed on August 25 and 26, 2011, respectively), re-writing the tool in another framework would make it easier for others to adopt the schema format and would potentially eliminate conversion bottlenecks that affect performance on large accounts.

EmailChemy. A proprietary Java application, EmailChemy can convert many proprietary and open mail formats stored as local files to open-format targets, such EML and MBOX files. It also includes a built-in IMAP server and can migrate converted messages into any IMAP-compliant server. Windows, Mac, and Linux/Solaris/Unix versions are available at http://www.weirdkid.com/products/emailchemy/

MailArchiva. A dual-licensed email journaling program, MailArchiva captures messages from many open source and proprietary mail transfer agents at the point of transmission or receipt, saving the messages to EML format, indexing them and providing access to them via a web interface. The developers provide an open source/community edition and an enterprise edition, with pricing based on the number of mailboxes being captured. The product is available at http://www.mailarchiva.com

Mailstore Server and Mailstore Home. Email archiving software marketed to small- and medium-size businesses, Mailstore Server allows organizations to capture a copy of email from any compatible IMAP server. Captured messages are kept in a central storage environment, where compression and single instance storage are used to optimize disk

usage. Mailstore records SHA1 hashes to impede tampering. The application will also find and integrate content from local folders (such as .pst files), and it includes e-discovery and record retention features. Emails may be accessed via Outlook or another IMAP-compatible client or device, and they can also be exported to several open formats (deepinvent Software GmgH 2011). Mailstore Home, which is free for non-commercial use, provides private individuals a method to back up all of their email accounts to a local computer or external drive, storing content in a proprietary format, while allowing export to system neutral formats such as EML. Information and downloads are available at http://www.mailstore.com/

Symantec Enterprise Vault. A hardware/software appliance targeted at large corporations and government agencies, Enterprise Vault is a proprietary tool for capturing and managing electronic messages and other documents to ensure compliance with legal and regulatory requirements. It can capture messages using several methods, typically by accessing journaling systems via the application programming interfaces for Microsoft Exchange, Lotus Domino and other SMTP compliant servers. It also includes a tool to pull messages from unstructured/dispersed storage environments, such as .pst files and .nsf archives. Once captured, messages are stored alongside other enterprise content in a database. Messages can be accessed directly from the user's client software program, such as Microsoft Outlook, using optional plug-ins. In their product literature, Symantec does not specific a storage format, but messages can be restored to their native format using the Outlook plug-in. By establishing rules, the system administrator can establish retention rules, enforce deletion policies and automate migration (Symantec Corporation 2011; Snyder 2007). Like most email archiving appliances, software packages or services, it makes extensive use of single instance storage and file compression.

Xena. A general-purpose file normalization tool developed by the National Archives of Australia, Xena uses the readpst library to convert Microsoft Outlook personal folders to mbox format. Xena is available at http://xena.sourceforge.net/

3.5. Email Search, Discovery, Access and Rendering Technologies

Because most email clients and servers are not intended to serve as long-term repositories, they typically include very weak search and discovery systems, even for messages in a single account. As the volume of messages increases, most client search tools slow to a crawl, particularly if messages are not mirrored to the client machine. Apple's Mail program and the specialist client The Bat! provide two exceptions. They include reasonably speedy filtering tools, which work against local copies of messages that may also be stored on the server. Some proprietary tools such as Rocketbox, Lookeen and Xobni can considerably improve the performance of in-client search and web-based search tools, such as those built in to Gmail and Hotmail.

Once messages have been migrated outside their native client/server architecture, they can be searched, retrieved and displayed by loading them back into another server or

client application. Therefore, repositories should keep the original email format or MBOX/EML files as an access copy, which can be imported into email clients as needed. Attachment display would depend on the file/software associations supported by the end user's client computer.

At present, several tools, including Hypermail and Aid4Mail, can convert messages to a static HTML format. In addition, a tool currently under development at Stanford University shows exciting potential for making preserved email useful. The Muse program, which can capture messages from several server environments to a local computer, also includes a search, browsing, visualization and analysis tool (Stanford University, Mobisocial Laboratory 2011).

If an institution chooses to convert email to an XML format for preservation purposes, individual messages would only be discoverable from the XML source after constructing a separate search index and access tool. Once this has been completed, the institution could integrate the source files into a preservation system (such as Fedora), index the files using a tool like Apache Lucene Solr, and make them discoverable using something like the Blacklight or VuFind applications (Villanova University 2011; University of Virginia 2011). In other words, implementing a merged search, discovery and access portal for email messages that have been preserved in XML would take considerable research and development work. Such work would be facilitated by collaboration with funding agencies and commercial entities that develop email software or design digital preservation services. Some web-based access tools have been developed by private parties; the open-source software used the former Alaska governor Sarah Palin's emails provides one example (The Sunlight Foundation 2011). In general, the development of search, discovery, retrieval and display tools for email should be a very high priority for the cultural heritage and digital preservation communities.

3.6. Email Preservation Case Studies

Using tools such as those described above, several institutions, including Harvard University, the State of Arizona Archives, the Smithsonian Institution Archives, and the Royal Library of Denmark are currently actively working to preserve email (Goethals and Gogel 2010; Schmitz Fuhrig and Adgent 2008; PeDALS Project 2010; Jensen 2011). Case studies of preservation actions undertaken at the Bodleian Library, Oxford, and the Medical Research Council, Glasgow, provide insight into the challenges and opportunities of preserving email.

3.6.1. Bodleian Library: Preservation for Cultural Heritage[5]

As the main research library at the University of Oxford, the Bodleian Library has been receiving a wide range of digital manuscripts, personal papers and organizational records. With funding from the Andrew W. Mellon Foundation, the Bodleian's futureArch project is developing methods to deal systematically with hybrid (analog and born digital) collections.

The approach that the Bodleian is taking towards email can best be thought of as an example of the 'sweeping up crumbs' approach described below, in which personal mailboxes are captured and migrated to a system-neutral format that utilizes the RFC standards. The Bodleian Library is developing processes to access and process 'dead' email collections, as well as those that continue to grow through active use.

In several instances, curatorial staff harvested emails from local folders or an email server, often with the help of the donor (or a colleague/heir). Records accessioned and converted include those from an email account of a former Member of Parliament/Member of the European Parliament (in Exchange/Outlook format), the professional email of an academic (in Compuserv 4.0 format), and files concerning a completed book project (in Gmail). In addition, the Bodleian is developing processes to manage incremental accessions. In one case, they are working with the email records of a small press (organized by publication); in another they are exploring options for capturing content from Lotus Notes/Domino mailboxes created by the staff members of an organization.

In each instance, curatorial staff identified the email as part of a broader records survey, recommending that the entire email account or a selected portion of it be transferred to the archives. In most instances, staff completed background research to locate the files and to understand their structure. Files were then copied to a portable hard drive or DVD, in whatever source format could be provided by the email client or server. Once the files had been received, futureArch staff identified software to read and/or migrate the files into the EML formats.

In the course of migrating these files, staff faced several challenges; none have so far proven to be insurmountable. In one case, over 200,000 messages were included in only two folders: sentmail and inbox. The migration program they used copied the messages without a problem, even though they were deposited in a single .pst file, split over five DVDs. While staff members are reasonably confident that the messages migrated correctly, verification could not be completed at the time of migration because the original file was so large that it could not be opened in versions of Outlook then accessible to the staff.

[5] This case study is based upon information presented by Susan Thomas at the Digital Preservation Coalition Briefing Day, 'Preserving Email: Directions and Perspectives,' July 29, 2011, and in subsequent correspondence with Ms. Thomas.

In another case, staff needed to migrate emails from the obsolete CompuServe format. This became a multi-step process. First they located and installed a copy of the CompuServe client application. The messages could not be viewed in the client because access required an active account. However, it proved fortunate that the client program could be located, because the migration tool that futureArch staff located (CS2Eduora) required the client software in order to operate. After finding an old version of Eudora on an abandoned website, staff members were able to view the emails, then migrate them into a format that could, in turn, be migrated to EML files.

In another instance, staff faced a significant challenge in segregating personal/private emails from professional correspondence (a donation requirement). In the end, futureArch staff used a forensics tool, FTK, to present a set of emails to the donor for review.

Finally, email stored in a Lotus Notes local folder (.nsf files) or on Lotus Domino servers is particularly problematic. Notes has a very poor end-user export facility. FutureArch staff wished to establish an IMAP connection to the Domino server, then use a client like Thunderbird to migrate the messages. Unfortunately, the Domino server was not configured to support IMAP connections, so staff negotiated with system administrators to see if selected mailboxes could be exported using server tools, which could then be migrated to EML, in a two-step migration process.

Oxford's experience illustrates four points:

- There are many variables to consider when capturing email. Each case requires different tools.
- The community needs characterization tools. For example, it would be highly desirable to complete audit verification, even if it was as simple as counting the number of emails and attachments before and after conversion.
- Researchers could benefit from visualization tools and client-like interfaces to make research use of email. In practical terms, files in EML format can be imported to most current clients, allowing access.
- Finally (and perhaps most importantly), successful email preservation projects are built on trusting relationships and professional competence. People are understandably nervous about donating semi-structured information like email to a repository. Curators must take those concerns seriously and provide effective mechanisms to address them.

Oxford is pursuing a practical and sustainable approach to dealing with email that might be accessioned as part of a broader records transfer. The techniques the Bodleian is developing will be applicable to many institutions, no matter their size or funding level, and will prepare them well for managing email within future digital repository settings.

3.6.2. Medical Research Council: Medium-Term Preservation[6]

The Social and Public Health Sciences Unit (SPHSU) of the Medical Research Council promotes human health by conducting research concerning the effects of environmental and social factors on people's physical and mental well-being. Based at the University of Glasgow, the unit directly employs about 100 people and collaborates with another 20 or so at any given time. Unit staff members maintain external research partnerships with others in the academic community and the National Health Service. As members of the MRC and their research partners conduct research, aggregate information, analyse datasets and write reports, they send and receive documentation concerning the research process via email.

Since 2007, the Unit has been using the dual-licenced program MailArchiva to mirror a copy of every sent and received message for the approximately 120 accounts managed by their qmail message transfer agent/server. The messages are written in EML format to an external store, located on a different physical machine than the sending/receiving server. MailArchiva keeps an index of the messages and generates a web-accessible discovery site, which includes filter and search features and which is integrated with existing authentication services. Using this interface, staff can view messages and optionally save them in EML format outside the system, from where they can be restored to the account or manipulated in other software.

The institution chose to implement this software for several reasons. By 2007, it was apparent that the volume of email on the sending/receiving server had outstripped available resources. With quotas in place, many users were writing email to local computers, losing important messages, or asking for restores from tape backup, even several years after they had deleted messages. While IT staff could accommodate most requests, they felt burdened by an inefficient storage and retrieval process. Against this background, the IT Systems Manager, Crawford Neilson, researched available options and consulted with his IT Strategies Committee, made up of unit researchers and other stakeholders. Since the unit director had an on-going interest in the topic of digital preservation and since members of the committee expressed a strong preference for treating email as a record of their research activities, it seemed reasonable to capture the entire stream of sent and received messages.

From Neilson's perspective, the system has proven successful. It was easily installed alongside their existing qmail server and since it runs on a different machine it causes little additional load on the server. While the software allows several configuration options, MRC has chosen to capture messages every 15 minutes, using the fetchmail utility. Over a period of three and a half years, it has captured in excess of 800,000 individual messages and has solved the previous storage problems. Since the community edition of MailArchiva uses single instance storage on a message level and compresses messages, the total

[6] This case study is based on information provided in an interview conducted by the author with IT Systems Manager Crawford Nielson on 11 August 2011.

storage volume for all email traffic sent by the 120 users over three years has amounted to less than 30 gigabytes.

The system was put in place with a few policy guidelines, which have been incorporated into the general IT policy that all employees are given upon hire. This policy simply states that each employee has a 2.5 GB limit on their personal account and that all sent and received messages will be captured to an external archive, which can be accessed at any time via a web browser. In addition, an employee's supervisors are provided with access to the account, and employees are told the software will continue to mirror their account for at least six months after they leave employment, after which their qmail account will be deleted. Employees can export messages from their archive at any time if they desire a personal copy. If necessary, system administrators can export a large volume of messages in EML or other formats, for import to other systems.

The software also includes the ability to set retention periods for particular types of records. Currently the SPHSU retains the complete archives created by the MailArchiva software for both current and former employees, as a record of corporate activities. At the time the system was implemented, the MRC planned to keep six years' worth of email. However, it is possible to save messages for longer time periods, and this policy may be reassessed and the period extended in view of the overall storage efficiencies that have been achieved. Presumably, the SPHSU could mark emails for very long-term or even permanent retention. One can easily imagine an institution using mail archiving software in this fashion, then creating a complementary policy that would allow for the transfer of email from selected individuals to a repository that provides long-term preservation services for emails holding cultural or historical value.

In short, the MRC is pursuing a policy of medium-term preservation and is doing so quite effectively. The software and hardware they are using stores messages in an open format that is based on the RFC 2822 and MIME standards. The software also provides a reasonably efficient search and discovery system. While the email archive provides end users a way to retrieve their own deleted messages, it also allows the institution to respond to potential legal and audit requests, while leaving the files in a preservation-ready format.

4. Recommended Actions

The case studies cited above demonstrate that institutions can use currently available tools to undertake concrete actions that will facilitate the long-term survival of email messages, accounts and systems. Given the array of different challenges and contexts that individual institutions will face, a wide range of solutions should be considered when looking at a local programme or project. Provided that all institutions should work toward the common goal of migrating and managing files in system-neutral formats, each situation will require services to be tailored to fit the circumstances. The author recommends that the following options be carefully considered when developing an email preservation programme.

4.1. For Institutions

There are three basic steps which institutions undertaking email preservation projects should undertake: defining policies, choosing appropriate tools, and implementing them in the light of local environmental factors and available resources.

As Maureen Pennock notes, institutions should start by defining email management and preservation policies in light of local conditions (Pennock 2006, p. 31). Email policies should outline: 1) an institutional commitment to email preservation, and specific actions that will be taken and supporting procedures; and 2) end-user expectations, responsibilities and rights regarding the access, use, privacy and control of email archives.

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 an entire lifecycle. User policies should be short, laying out acceptable uses, etiquette, effective communication styles, 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 application (such as those discussed in section 3.4), so that users benefit from a technical framework that automatically segregates archived messages to an external, server-based repository. Records management or archival staff should provide users with basic training information to help ensure policy effectiveness.

Once appropriate policies have been defined, institutions should select and implement appropriate tools that support the policies, drawing on the information provided in section 3 of this report and in other resources such as those cited in section 7. Implementation must be undertaken in collaboration with email users, records managers, curators, archivists, IT managers and administrators. In particular, institutions should be careful not to inadvertently impose preservation-hostile configurations on users. This can be accomplished by providing adequate storage space and avoiding auto-deletion settings. In general, five preservation strategies suggest themselves; one or more of these may be applicable depending on local circumstances:

1. **Sweeping Up Crumbs:** The ‘sweeping up crumbs’ or whole-account approach refers to harvesting email found on a user’s computer or account, working directly with her or her heirs. In practical terms, many institutions that have not previously had an email preservation policy or had influence over the design of record keeping systems will need to use this approach. In some cases, the work may best be accomplished by working in conjunction with IT staff to secure a copy of the email in its native format. Once native files have been secured, they can be migrated using the tools discussed above, then stored in a trusted digital repository. Section 3.6.1 provides an example of this approach used effectively at Oxford University. Toolslike Aid4Mail and Emailchemy could be used to implement similar projects.

2. **Nurturing and Harvesting:** The 'nurturing and harvesting' approach is an enhanced version of the 'sweeping up crumbs' approach, wherein an institution offers guidance and assistance to email users during the lifetime of systems, helping them to ensure that critical records are retained in system-neutral formats, then using email migration software to capture and preserve records either as they are created or at the end of a user's lifetime. The University of Michigan implements a version of this approach by providing users a designated 'archives' box in their email client, to which they can drag and drop emails, which are then co-managed by email users and archival staff members (Shallcross 2011).
3. **Capturing Carbon:** The 'capturing carbon' or whole system approach implements email archiving software to capture an entire email ecosystem or a portion of that ecosystem to an external email storage environment. Optionally, rules can be applied either at time of capture or disposition; such rules would have retention periods associated with senders, recipient, dates sent or received, keywords, or folder titles. Ideally, records will be written in a system-neutral format, allowing for the integration of records into a trusted digital repository. The Medical Research Council case study cited in section 3.6.2 provides an example of this approach. In the United States, one effort to preserve an academic listserv also used this general approach (Schmidt 2011).
4. **Tagging and Bagging:** the 'tagging and bagging' or message-by-message approach implements an enterprise-level electronic records management system or informal classification system to which email messages might be declared as 'records,' then saved alongside non-email records relating to the same function or activity. As far as I am aware, no published literature testifies to the effectiveness of this approach, either in ensuring short-term compliance with law or policy or in preserving email for the long-term. However, one suggested implementation has been proposed (Wagner, et al., 2008a; Wagner, et al., 2008b). Institutions that already use an EDRMS system or are considering one may wish to pursue this option, but care should be taken to ensure the system is configured so that it supports and takes advantage of users' normal email work habits. Automatic classification services may prove useful in making this approach feasible.
5. **Personal Archives Service:** This approach refers to a prospective service, in which a non-profit agency or for-profit corporation provides a service to harvest email messages from a user's account to a trusted, cloud-based and replicated environment, using an open and system-neutral format. Based on ideas first suggested by Eric Freeman, David Gelernter and David Bearman, such a service would operate similarly to services such as Carbonite or CrashPlan, but would offer people the opportunity to donate their email to an archives, manuscript repository or other service provider; one such service is currently being explored by the author as a research project (Freeman and Gelernter 1996; Bearman and Hedstrom 2000; Prom 2011). One American university is piloting a similar service with selected graduate students (University of North Carolina School of Information and Library

Science 2011). In addition, the Muse tool cited in section 3.5 could play a role in the development of a more systematic service of this type.

While the last option is speculative, the tools that would be necessary to build such a service currently exist. They are drawn from options one to four, which can be achieved using current technologies.

4.2. For Individuals

Regardless of the services that an institution provides, many people will wish to preserve their email and other electronic communications so that they can use the messages as a personal memory bank or even so they can donate them to a repository at a later time. Such individuals can take several simple steps to facilitate the long-term preservation of sent and received email messages.

First and foremost, they should familiarize themselves with the operation of the email services that they use, as well as the clients that they use to operate them. Several resources provide advice that is intended to help users manage their email within a server, client and local storage infrastructure (How-To Geek 2007; C. Prom 2010; Schmitz Fuhrig 2011). Once email users understand their account structures and have configured them to manage email more effectively, they might use simple backup tools, such as the free MailStore Home application or the Apple Time Machine backup tool, to save a copy of their email records in a secure location, isolated from the copies used directly by the application. MailStore and Time Machine are particularly attractive options because they can be used to take snapshots of accounts over time. One disadvantage of MailStore Home is that it must be manually run each time the user wishes to make a backup. Time Machine, on the other hand can only play a role in a preservation strategy if the person using it has configured an email client program to download copies of the messages to the local machine.

Individuals might also use a cloud-based service to backup their email, but they should carefully consider the terms of service before subscribing. Online backup services such as Carbonite and Mozy, available at no or modest cost to personal users, automatically back up a copy of everything in a personal computer's data folders, including local copies of email messages, kept in whatever format the user's client machine stores users. Similar services for social media and cloud-based email services such as Gmail have recently been made available or are under development (Rafe 2011; Nuffly.com 2011). While such options may offer a short-term backup plan and have the advantage of copying all a user's files, including those communications sent with tools other than email, they carry some risks. For example, the terms of service for Mozy explicitly state that the service provider has no obligation to return data to the user in the event of account termination or business failure. In any case, it is difficult to say how one would get data out of the backup system and into a preservation-ready format. Therefore, users of cloud-based backup services may wish to use a service that also allows them to store data on a local backup device or on a friend's computer (such as CrashPlan), or they may wish to use online backup as a

supplement to local backup. In any case, it is very important to understand the details of how the chosen backup or archive system works with regard to email. In particular, most services will only provide the ability to recreate the last system state, not to restore emails deleted at a previous time.

4.3. For the Digital Preservation Community

Although many concrete preservation actions can be taken by organizations and by individuals, the digital preservation community could benefit from a more complete understanding of how email can be preserved and from the development of additional tools that capture, store and provide access to preserved email. Records managers, archivists, curators, librarians, resource allocators, grant-giving agencies and others interested in preserving email can take several actions to develop the next generation of email preservation services:

- digital preservation leaders can advocate for the necessity of email preservation as a cultural and public good;
- archivists, records managers and IT managers can undertake research work and case studies (in regards to user behaviours, legal issues and technical approaches to preserving email), with a stronger emphasis on peer-reviewed studies;
- organizational leaders can write and publicize simple, specific and concise email management policies and advice documents that take advantage of end users' email preferences and behaviours;
- computer programmers can develop access tools, so that messages stored in XML formats may be browsed, searched, displayed and visualized;
- grant-giving agencies can encourage the growth of a research and development agenda regarding email preservation; and
- private companies and libraries can provide trusted personal archiving services for current record creators and potential donors of email and other personal digital archives.

5. Conclusion

During the run up to the Second Gulf War, in late 2002 and early 2003, IT staff in the Executive Office of the President of the United States replaced their Lotus Domino servers with Microsoft Exchange servers, possibly losing over 22 million emails in the process (Gewirtz 2007; Gewirtz 2009). On 19 November 2009, over 1,000 emails and 3,000 other private documents that had been stolen from the University of East Anglia's Climatic Research Unit were uploaded to a web server in Russia and immediately mirrored across the Internet, affecting scientific and political debates over an issue of great public prominence. On 1 September 2011, Der Spiegel revealed that a corpus of 250,000 unredacted US State Department emails in the possession of Wikileaks had been made

available on the Internet, possibly imperilling diplomatic negotiations and the personal safety of individuals worldwide (Leigh 2011; Mackey, Harris, Somaiya and Kulish 2011).

These incidents demonstrate that while email can hold extraordinary public interest and historical value, it is also very fragile and susceptible to loss or abuse. By using the methods and software discussed in this report and by developing tools and services that support additional preservation options, we in the cultural heritage community can make the preservation of trusted email records a systematic part of our everyday operations. In doing so, we will save records as rich as the paper-based letters and memoranda that now fill archives and manuscript repositories. Those who write the history of our era will thank us for our efforts.

6. Glossary and Acronyms

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 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.

Exchange Server: A proprietary application developed and licensed by Microsoft Corporation, providing server-based email, calendar, contact and task management features. Exchange servers are typically used in conjunction with Microsoft Outlook or the Outlook Express web agent. Exchange servers use a proprietary storage format and messages sent using Exchange typically include extensive changes to the header of the file. Calendar entries, 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 an IMAP-aware client application.

Internet Message Format (IMF): A 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.

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. 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 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.

Milter: an extension for the popular Postfix and Sendmail mail transfer agents, allowing for the identification, sorting and filtering of messages while they are in transit from the sending to the receiving server. Although milters are typically used to identify and quarantine spam and viruses, they can also be used to identify and filter messages for capture to an external email store.

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 needed.

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 typically possible to direct the server to maintain the message by setting a configuration directive in the email client.

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.

Single Instance Storage: a method by which a computer system keeps and points to one copy of a message, document or other data, even though the data is shared among multiple users or accounts, thereby de-duplicating information and saving storage space.

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.

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.

7. Further Reading

The following resources are particularly useful. Complete citations for each item can be found in the References section.

David Bearman's 1994 article, 'Managing Electronic Mail' outlines the major social, technical and legal issues that any email preservation project needs to address. It is particularly useful in suggesting ways that systems designs can support the effective implementation of policies. Tools developed since 1994 make Bearman's approach much more practical now, than at the time of publication.

Maureen Pennock's 2006 'Curating Emails' remains essential reading, reviewing the major challenges to email preservation, and summarizing some prospective approaches. She places particular emphasis on the need to manage email effectively during its period of creation and active use so that is left in a preservation-ready format. The work also outlines the major conceptual approaches that can be used to preserve email, with somewhat less description of particular tools or services.

Richard Cox's 2008 book chapter on email preservation reviews the history of attempts that the archival profession has made in preserving email messages and their content, suggesting that the best approaches will understand and preserve them as the organic outcome of our professional and personal lives. Cox suggests that those wishing to

preserve email draw on concepts and procedures from both the records management and manuscript archives traditions, but the chapter contains relatively little direct implementation advice.

Gareth Knight's 2009 report on email migration tools, completed for the InSPECT project, includes a description and analysis of the structure of an email message, identifying 14 properties of the message header and 50 properties of the message body that must be maintained during migration if an email is to be considered authentic and complete. The report also outlines a procedure for testing whether particular email migration tools preserve those properties and applies that procedure to three specific tools.

Andrea Goethals' and Wendy Gogel's 2010 iPres conference paper, outlining Harvard University's plan to support email preservation within a local digital repository, provides an example of attempts to integrate email preservation into existing library and archival services. Reviewing major challenges and discussing the formation of the project team and implementation decisions, the authors outline the process that led them to choose an XML format for preservation, noting relevant administrative issues and outlining prospective deposit and processing workflows.

8. References

Anderson, C., 2011. Help Create an Email Charter! TEDChris: The Untweetable. Available at: http://tedchris.posterous.com/help-create-an-email-charter [Accessed July 15, 2011].

Ashenfelder, M., 2011a. Personal Archiving in the Cloud. The Signal: Digital Preservation. Available at: http://blogs.loc.gov/digitalpreservation/2011/06/personal-archiving-in-the-cloud/ [Accessed June 9, 2011].

Ashenfelder, M., 2011b. When I Go Away: Getting Your Digital Affairs in Order. The Signal: Digital Preservation. Available at: http://blogs.loc.gov/digitalpreservation/2011/07/when-i-go-away-getting-your-digital-affairs-in-order/ [Accessed July 3, 2011].

Bailey, S., 2011a. Paying Lip Service to the User. Records Management Futurewatch. Available at: http://rmfuturewatch.blogspot.com/2011/06/paying-lip-service-to-user.html [Accessed June 10, 2011].

Bailey, S., 2011b. Email Management: Fifteen Wasted Years and Counting. Available at: http://e-records.chrisprom.com/?p=2284 [Accessed June 20, 2011].

Baron, J., 2010. The Future of Email Preservation. Available at: www.archives.gov/records-mgmt/pdf/baron-raco2010.pdf [Accessed June 20, 2011].

Barzun, J. and Graff, H.F., 1992. The Modern Researcher 5th edn., Fort Worth, Texas: Harcourt Brace Jovanovich College Publishers.

Beagrie, N., 2005. Plenty of Room at the Bottom? Personal Digital Libraries and Collections. D-Lib Magazine, 11(06). Available at:

http://www.dlib.org/dlib/june05/beagrie/06beagrie.html [Accessed August 30, 2011].

Bearman, D., 1993. The Implications of Armstrong v. Executive of the President for the Archival Management of Electronic Records. American Archivist, 56(4), pp. 674–689.

Bearman, D., 1994. Managing Electronic Mail. Archives and Manuscripts, 22(1), pp. 28–50.

Bearman, D. and Hedstrom, M., 2000. Reinventing Archives for Electronic Records: Alternative Service Delivery Options. In American Archival Studies: Readings in Theory and Practice. Society of American Archivists, pp. 549–567.

Beebe, D., 2008. Enterprise Vault and Discovery Accelerator: Email Archiving and Discovery Solution Implementation and the Legal Landscape. Denver Colorado: Regis University. Available at: http://adr.coalliance.org/codr/fez/view/codr:288 [Accessed June 3, 2011].

Billsus, D. and Hilbert, D., Seamless Electronic Mail Capture With User Awareness And Consent. Available at: http://www.google.com/patents?hl=en&lr=&vid=USPATAPP11620850&id=BL2iAAAAEBAJ&oi=fnd&dq=%22email+archiving%22&printsec=abstract#v=onepage&q=%22email%20archiving%22&f=false [Accessed June 11, 2011].

BlogForever Project, 2011. BlogForever Project Website. Available at: http://blogforever.eu/ [Accessed July 20, 2011].

Boudrez, F., 2006. Filing and Archiving Email, Antwerp: Expertisecentrum DAVID vzw. Available at: http://www.expertisecentrumdavid.be/docs/filingArchiving_email.pdf [Accessed June 3, 2011].

Boudrez, F. and Van Den Eynde, S., 2002. DAVID: Archiving Email, Antwerp: City of Antwerp Archives. Available at: http://www.expertisecentrumdavid.be/davidproject/teksten/Rapporten/Report4.pdf [Accessed June 3, 2011].

Brodkin, J., 2011. The MIME Guys: How Two Internet Gurus Changed E-mail Forever. Network World. Available at: http://www.networkworld.com/news/2011/020111-mime-internet-email.html?page=1 [Accessed August 31, 2011].

Brogan, M. and Vreugdenburg, S., 2008. 'You've Got Mail': Accountability and End User Attitudes to Email Management. In Proceedings of the 4th International Conference on e-Government. RMIT University Melbourne Australia, pp. 63–69.

Buckles, G., 2011. Custodial Email Preservation – Email Infestation. eDiscovery Journal, (March 2011). Available at: http://ediscoveryjournal.com/2011/03/custodial-email-preservation-%E2%80%93-email-infestation/ [Accessed August 30, 2011].

California Digital Library, 2011. Web Archiving Service. Available at: http://webarchives.cdlib.org/ [Accessed July 18, 2011].

Carden, M., 2011. Email Message sent to Christopher Prom on July 19, 2011. Copy available from recipient upon request.

Carroll, E. and Romano, J., 2011. Your Digital Afterlife: When Facebook, Flickr and Twitter Are Your Estate, What's Your Legacy?, Berkeley, CA: New Riders.

Cavender, A., 2010. Backing Up a Campus Email Account: GMail, iCal, and a Desktop Application. ProfHacker – The Chronicle of Higher Education. Available at: http://chronicle.com/blogPost/Backing-Up-a-Campus-Email-A/25992/ [Accessed July 16, 2011].

Charlesworth, A., 2009. Digital Lives: Legal and Ethical Issues: A Discussion Paper, London: The British Library. Available at: http://britishlibrary.typepad.co.uk/files/digital-lives-legal-ethical.pdf [Accessed June 10, 2011].

Chatelain, J.-Luc and Garrie, D.B., 2007. The Good, the Bad and the Ugly of Electronic Archiving: An Essay on the State of Enterprise Information Management. Journal of Legal Technology Risk Management, 2(1), pp. 90–97.

Cole, R. and Eklund, P., 1999. Analyzing an Email collection using formal concept analysis. In Principles of Data Mining and Knowledge Discovery. Proceedings of the Third European Conference, PKDD'99. (Lecture Notes in Artificial Intelligence Vol. 1704). Berlin, Germany, pp. 309–15.

comScore, Inc., 2011. Email Evolution: Web-based Email Shows Signs of Decline in the U.S. While Mobile Email Usage on the Rise. Available at: http://www.comscore.com/Press_Events/Press_Releases/2011/1/Web-based_Email_Shows_Signs_of_Decline_in_the_U.S._While_Mobile_Email_Usage_on_the_Rise [Accessed July 17, 2011].

Consultative Committee for Space Data Systems, 2002. Reference Model for an Open Archival Information System (OAIS), Available at: http://public.ccsds.org/publications/archive/650x0b1.pdf [Accessed June 10, 2011].

Cox, R.J., 2008. Chapter 7: Electronic Mail and Personal Recordkeeping. In Personal Archives and a New Archival Calling: Readings, Reflections and Ruminations. Duluth, Minnesota: Litwin Books, pp. 201–42.

Crispin, M., 2003. RFC 3501 – Internet Message Access Protocol – Version 4, Revision 1. Available at: http://tools.ietf.org/html/rfc3501 [Accessed August 30, 2011].

Crook, C., 2010. Climategate and the Big Green Lie – Politics – The Atlantic. Available at: http://www.theatlantic.com/politics/archive/2010/07/climategate-and-the-big-green-lie/59709/ [Accessed July 22, 2010].

Croxall, B., 2010. Backing up Your Social Network v2.0. Profhacker – The Chronicle of Higher Education. Available at: http://chronicle.com/blogs/profhacker/backing-up-your-social-network-v20/26890 [Accessed June 6, 2011].

Dale, R. and Ambacher, B. eds., 2007. Trustworthy Repositories Audit & Certification (TRAC) : Criteria and Checklist, Chicago, Illinois: CRL.

Digital Preservation Testbed, 2003. From Digital Volatility to Digital Permanence: Preserving Email, The Hague: Dutch National Archives. Available at: http://en.nationaalarchief.nl/sites/default/files/docs/kennisbank/volatility-permanence-email-en.pdf.

DRAMBORA Project, Digital Repository Audit Method and Risk Assessment. Available at: http://www.repositoryaudit.eu/ [Accessed July 22, 2010].

Enneking, N., 1998. Managing Email: Working Toward an Effective Solution. Records Management Quarterly, 32(3), p. 24.

ExactByte, LLC, 2011. Tweetymail Website. Available at: http://tweetymail.com/ [Accessed July 16, 2011].

ExactTarget, 2010. Is email on the decline? Event360.com. Available at: http://www.event360.com/blog/is-email-on-the-decline/ [Accessed July 17, 2011].

Fallows, J., 2011. Hacked! The Atlantic. Available at: http://www.theatlantic.com/magazine/archive/2011/10/hacked/8673/# [Accessed November 11, 2011].

Foggo, G., Grosso, S., Harrison, B. and Rodriguez-Barrera, J.V., 2007. Comparing E-Discovery 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). Available at: http://www.mcmillan.ca/Files/BHarrison_ComparingE-Discoveryintheunitedstates.pdf [Accessed June 3, 2011].

Freeman, E. and Gelernter, D., 1996. Lifestreams: a Storage Model For Personal Data. ACM SIGMOD Record, 25(1), pp. 80–86.

Gewirtz, D., 2007. Where Have All the Emails Gone?, Palm Bay, Fla.: Zatz. Available at: http://www.worldcat.org/title/where-have-all-the-emails-gone/oclc/227004599 [Accessed July 10, 2011].

Gewirtz, D., 2009. Some Bush-era Emails Restored, Many Still Lost to History. Anderson Cooper 360 Blogs. Available at: http://ac360.blogs.cnn.com/2009/12/14/some-bush-era-emails-restored-many-still-lost-to-history/ [Accessed July 7, 2011].

Goethals, A. and Gogel, W., 2010. Reshaping the Repository: The Challenge of Email Archiving. In 7th International Conference on Preservation of Digital Objects (iPRES2010*)*. Vienna, Austria. Available at: http://www.ifs.tuwien.ac.at/dp/ipres2010/schedule.html [Accessed June 2, 2011].

Gorton, D., Murthy, U., Vemuri, S. and Perez-Quinones, M.A., 2007. Email-Set Visualization: Facilitating Re-Finding in Email Archives. Available at: http://eprints.cs.vt.edu/archive/00000956/ [Accessed June 9, 2011].

Green, M., Soy, S., Gunn, S. and Galloway, P., 2002. Coming to TERM: Designing the Texas Email Repository Model. D-Lib Magazine, 8(9). Available at: http://www.dlib.org/dlib/september02/galloway/09galloway.html [Accessed June 2, 2011].

Gregory, A., 2010. 13 Tools to Back Up Your Social Media Content. SitePoint. Available at: http://www.sitepoint.com/backup-social-media-profiles/ [Accessed July 16, 2011].

Gruenspecht, J., 2011. 'Reasonable' Grand Jury Subpoenas: Asking for Information in the Age of Big Data. Harvard Journal of Law & Technology, 24(1), pp. 543–62.

The Guardian, 2011. Hacked Climate Science Emails. Available at: http://www.guardian.co.uk/environment/hacked-climate-science-emails [Accessed November 27, 2011].

Guy, M., 2011. Preserving your Emails. JISC Beginner's Guide to Digital Preservation. Available at: http://blogs.ukoln.ac.uk/jisc-bgdp/2011/03/02/preserving-your-emails/ [Accessed June 3, 2011].

Hall, E., 2005. RFC 4155 – The Application/Mbox Media Type. Available at: http://tools.ietf.org/html/rfc4155 [Accessed August 31, 2011].

Harbaugh, L.G., 2010. Sorting through Email Archiving Tools. Network World. Available at: http://www.networkworld.com/reviews/2010/101110-email-archiving-test.html [Accessed July 4, 2011].

Harbaugh, L.G., 2011. Iron Mountain Wins Email Archiving Test. Network World. Available at: http://www.networkworld.com/reviews/2011/022111-email-archiving-test.html [Accessed July 4, 2011].

Hill, B.W., 2011. The Forrester Wave™: Message Archiving Software, Q1 2011 – Forrester Research. Available at: http://www.forrester.com/rb/Research/wave%26trade%3B_message_archiving_software,_q1_2011/q/id/53276/t/2 [Accessed July 4, 2011].

How-To Geek, 2007. Force Outlook 2007 to Download Complete IMAP Items. *How-To Geek*. Available at: http://www.howtogeek.com/howto/microsoft-office/force-outlook-2007-to-download-complete-imap-items/ [Accessed July 16, 2011].

Howard, S., 2011. Why Preserving Email is Harder than it Sounds: Steven Howard | Practical E-Records. In DPC Briefing: Preserving Email: Directions and Perspective. Available at: http://e-records.chrisprom.com/?p=2192 [Accessed August 30, 2011].

Hunter, P., 2007. Email Meets Enron to Bring Lawyers Down on Big Corporations. Computer Fraud & Security, 2007(5), pp. 18–20.

Hyry, T. and Onuf, R., 1997. The Personality of Electronic Records: The Impact of New Information Technology on Personal Papers. Archival Issues, 22(1), pp. 37–44.

IBM Corporation, 2009. Demo: Lotus Domino Replication Basics. Available at: http://www-10.lotus.com/ldd/dominowiki.nsf/dx/domino-replication-basics [Accessed September 19, 2011].

Internet Archive, 2011. Archive-It.org. Available at: http://www.archive-it.org/ [Accessed July 18, 2011].

Internet Memory Foundation, 2011. Archivethe.net. Available at: http://internetmemory.org/en/index.php/projects/atn [Accessed July 18, 2011].

InterPARES Project, Project Website. Available at: http://www.interpares.org/ [Accessed July 30, 2011].

Jackson, T., 2009. The E-mail Optimisation Toolkit, London: ARK Group.

Jensen, C., 2010. Letting scholars auto-archive (a pragmatic solution to the acquisition and archiving of born-digital material at The Royal Library of Denmark). Presentation to

4th Conference of LIBER Manuscript Librarians Group: "Meeting with manuscripts, today and tomorrow", 27 May, 2010, Rome Italy. Copy available from author.

Juhnke, D., 2003. Electronic Discovery in 2010. Information Management Journal, 37, pp. 34–42.

Klensin, J., 2008. RFC 5321 – Simple Mail Transfer Protocol. Internet Engineering Task Force. Available at: http://tools.ietf.org/html/rfc5321 [Accessed July 20, 2011].

Klyne, G., 2003. An XML Format for Mail and Other Messages. Available at: http://www.ninebynine.org/IETF/Messaging/draft-klyne-message-xml-00.txt [Accessed June 2, 2011].

Knight, G., 2009. InSPECT – Email Testing Report. InSPECT: Investigating Significant Properties of Electronic Content. Available at: http://www.significantproperties.org.uk/email-testingreport.html [Accessed June 17, 2011].

Lappin, J., 2011. Preserving E-mail – Records Management Perspectives. Thinking Records. Available at: http://thinkingrecords.co.uk/2011/08/11/preserving-e-mail-records-management-perspectives/ [Accessed August 18, 2011].

Leigh, D., 2011. *Wikileaks: Inside Julian Assange's War on Secrecy* 1st edn., New York: Public Affairs.

Levsen, L., 2009. Comprehensive Network Analysis Shows Climategate Likely to Be a Leak. Watts Up With That? Available at: http://wattsupwiththat.com/2009/12/07/comprhensive-network-analysis-shows-climategate-likely-to-be-a-leak/ [Accessed July 18, 2011].

Li, Y. and Somayaji, A., 2005. Securing Email Archives through User Modeling. In Proceedings of 21st Annual Computer Security Applications Conference. pp. 547–556

Library of Congress, 2010. Extractor Tool Helps Preserves Microsoft Outlook Emails and Attachments. Library of Congress Digital Preservation: News and Events. Available at: http://www.digitalpreservation.gov/news/2010/20100924news_article_pedals_email_tool.html [Accessed May 25, 2011].

Library of Congress, 2011. NDIIPP Partner Tool and Service Inventory. Available at: http://www.digitalpreservation.gov/partners/resources/tools/index.html [Accessed August 31, 2011].

Lorenz, C., 2007. The Death of E-mail. Slate Magazine. Available at: http://www.slate.com/id/2177969/ [Accessed July 17, 2011].

Mackenzie, M.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.

Mackey, R., Harris, J., Somaiya, R. and Kulish, N., 2011. All Leaked U.S. Cables Were Made Available Online as WikiLeaks Splintered. The Lede: New York Times. Available at: http://thelede.blogs.nytimes.com/2011/09/01/all-leaked-u-s-cables-were-made-available-online-as-wikileaks-splintered/ [Accessed September 20, 2011].

Maher, W., 1992. Chaos and the Nature of Archival Systems. Presented at the Society of American Archivists 56th Annual Meeting Montreal, Quebec, Canada, September 15, 1992. Available at: http://www.library.illinois.edu/archives/workpap/chaosshort.pdf [Accessed July 10, 2011].

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

Marshall, C.C., 2008a. Rethinking Personal Digital Archiving Part 1: Four Challenges from the Field. D-Lib Magazine, 14(4). Available at: http://www.dlib.org/dlib/march08/marshall/03marshall-pt1.html [Accessed July 17, 2011].

Marshall, C.C., 2008b. Rethinking Personal Digital Archiving, Part 2: Implications for Services, Applications, Institutions. D-Lib Magazine, 14(3/4). Available at: http://www.dlib.org/dlib/march08/marshall/03marshall-pt2.html [Accessed August 31, 2011].

Meyer, C., 2009. Evolutions in Email Style and Usage. In Science and Technology for Humanity (TIC-STH), 2009 IEEE Toronto International Conference. pp. 609–612.

Microsoft Corporation, 2011. Outlook 2010 MAPI Reference. Available at: http://msdn.microsoft.com/en-us/library/cc765775.aspx [Accessed August 18, 2011].

Milicchio, F. and Gehrke, W., 2007. Electronic Mail. In Distributed Services with OpenAFS. Springer Berlin Heidelberg, pp. 237–262. Available at: http://dx.doi.org/10.1007/978-3-540-36634-8_9 [Accessed July 13, 2011].

Ministry of Justice, 2011. Civil Procedure Rules Homepage, 57th Update. Available at: http://www.justice.gov.uk/guidance/courts-and-tribunals/courts/procedure-rules/civil/index.htm [Accessed September 19, 2011].

Minor, D., 2008. Mail Account XML Schema: How Internet Messages Can Be Stored as XML. Available at: http://siarchives.si.edu/cerp/David_Minor_CERP_symp.pdf [Accessed May 25, 2011].

Moidu, S., 2009. Share Photos and Videos Anywhere You Have Email. Facebook. Available at: http://www.facebook.com/blog.php?post=109768117130 [Accessed July 17, 2011].

Morris, E., 2011. Did My Brother Invent E-Mail With Tom Van Vleck? (Part Three). NYTimes.com. Available at: http://opinionator.blogs.nytimes.com/2011/06/21/did-my-brother-invent-e-mail-with-tom-van-vleck-part-three/ [Accessed June 23, 2011].

Multicians.org, 2011. Multics History. Available at: http://www.multicians.org/history.html [Accessed July 6, 2011].

Myers, J., 1996. RFC 1939 – Post Office Protocol – Version 3. Internet Engineering Task Force. Available at: http://tools.ietf.org/html/rfc1939 [Accessed July 24, 2011].

National Library of Australia, 2011. PADI – Digital Preservation Tools. Available at: http://www.nla.gov.au/padi/topics/535.html [Accessed August 31, 2011].

North Carolina Office of Archives and History, 2009. Preservation of Electronic Mail Collaboration Initiative: Technical Resources. Available at: http://www.records.ncdcr.gov/emailpreservation/ [Accessed August 31, 2011].

Nuffly.com, 2011. Services – Nuffly.com. Available at: http://nuffly.com/services [Accessed July 16, 2011].

Open Planets Foundation, 2011. Community Website. Available at: http://www.openplanetsfoundation.org/ [Accessed August 31, 2011].

Paquet, L., 2000. Appraisal, Acquisition and Control of Personal Electronic Records: From Myth to Reality. Archives and Manuscripts, 20(2), pp. 71– 91.

Partridge, C., 2008. The Technical Development of Internet Email. Annals of the History of Computing, IEEE, 30(2), pp. 3–29.

PeDALS Project, 2010. PeDALS Email Extractor Software | Sourceforge, Available at: http://sourceforge.net/projects/pedalsemailextr/ [Accessed July 16, 2011].

Pennock, M., 2006. Curating E-Mails: A Life-cycle Approach to the Management and Preservation of E-mail Messages. In DCC Digital Curation Manual. Digital Curation Centre. Available at: http://www.dcc.ac.uk/sites/default/files/documents/resource/curation-manual/chapters/curating-e-mails/curating-e-mails.pdf [Accessed June 2, 2011].

Perer, A., Shneiderman, B. and Oard, D.W., 2006. Using Rhythms of Relationships to Understand E-mail Archives. Journal of the American Society for Information Science and Technology, 57(14), pp. 1936–1948.

Perry, T.S., 1992. Electronic Mail - Forces for Social Change. Spectrum, IEEE, 29(10), pp. 30–32.

Pinguelo, F.M. and Gonnello, F.J., 2010. Zubulake Revisited: Ineffective Lit Holds and Sloppiness Lead To Wheel of Sanctions. e-Lessons Learned. Available at: http://ellblog.com/?p=2009#more-2009 [Accessed July 22, 2011].

Potter, M., 2002. XML For Digital Preservation: XML Implementation Options for E-Mails. Available at: www.imaginar.org/dppd/DPPD/183%20pp%20XML.pdf [Accessed June 2, 2011].

Prom, C., 2010. Email Management and Preservation Guidelines. Practical E-Records. Available at: http://e-records.chrisprom.com/?page_id=1301 [Accessed July 16, 2011].

Prom, C., 2010. Software/Tools. Practical E-Records. Available at: http://e-records.chrisprom.com/?page_id=175 [Accessed August 31, 2011].

Prom, C., 2011. About the iKive Personal Archives Service. Available at: http://www.ikive.com/about/ [Accessed September 20, 2011].

Pukkawanna, S., Visootfiviseth, V. and Pongpaibool, P., 2006. Classification of Web-based Email Traffic in Thailand. In International Symposium on Communications and Information Technologies, 2006. ISCIT '06. pp. 440–445.

Radicati Group, Inc., 2011a. Email Statistics Report, 2011–2015, Executive Summary. Available at: http://www.radicati.com/?p=7261 [Accessed July 17, 2011].

Radicati Group, Inc., 2011b. Email Archiving Market, 2011–2015: Executive Summary. Available at: http://www.radicati.com/?p=7359 [Accessed July 21, 2011].

Rafe, N., 2011. Backupify is More than a Backup Service. Rafe's Radar | CNET News. Available at: http://news.cnet.com/8301-19882_3-20030614-250.html [Accessed June 6, 2011].

Resnick ed., 2008. RFC 5322 – Internet Message Format. Available at: http://tools.ietf.org/html/rfc5322 [Accessed June 7, 2011].

Ritchel, M., 2010. E-Mail's Big Demographic Split. New York Times | Bits Blog. Available at: http://bits.blogs.nytimes.com/2010/12/21/e-mails-big-demographic-split/ [Accessed July 17, 2011].

Schmidt, L., 2011. Preserving the H-Net Email Lists: A Case Study in Trusted Digital Repository Assessment. American Archivist, 74(1), pp. 257–296.

Schmitz Fuhrig, L., 2011. Guidelines for Managing Your Work and Personal Email. The Atlantic – Technology. Available at: http://www.theatlantic.com/technology/archive/2011/04/guidelines-for-managing-your-work-and-personal-email/237961/ [Accessed April 29, 2011].

Schmitz Fuhrig, L. and Adgent, N., 2008. Preserving Historical Correspondence: Email Preservation Progress and Future Directions. Available at: http://siarchives.si.edu/cerp/CERP_EMCAP_symp.pdf [Accessed June 10, 2011].

Scholtes, J., 2006a. A View on Email Management: Balancing Multiple Interests and Realities of the Workplace. KMWorld, pp. S6–7.

Scholtes, J., 2006b. Efficient and Cost-Effective Email Management with XML. KMWorld, p. S16.

Shallcross, M., 2011. The MeMail Project. Practical E-Records. Available at: http://e-records.chrisprom.com/?p=1965 [Accessed September 1, 2011].

Smallwood, R., 2008. Taming the Email Tiger: Email Management for Compliance, Governance, and Litigation Readiness: A Management Guide Original edn, New Orleans, LA: Bacchus Business Books.

Smithsonian Institution Archives, 2008. The Collaborative Electronic Records Project. Available at: http://siarchives.si.edu/cerp/ [Accessed May 25, 2011].

Snyder, J., 2007. Symantec's Enterprise Vault finds what you're looking for. Available at: http://www.networkworld.com/reviews/2007/061107-symantec-test.html [Accessed August 18, 2011].

deepinvent Software GmgH, 2011. MailStore Home, Available at: http://www.mailstore.com/en/mailstore-home.aspx [Accessed July 16, 2011].

Srinivasan, A. and Baone, G., 2008. Classification Challenges in Email Archiving. In C.-C. Chan, J. Grzymala-Busse, & W. Ziarko, eds. Rough Sets and Current Trends in Computing. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, pp. 508–519. Available at: http://dx.doi.org/10.1007/978-3-540-88425-5_53.

Stanford University, Mobisocial Laboratory, 2011. Muse Home Page. Available at: http://mobisocial.stanford.edu/muse/ [Accessed September 19, 2011].

Sudarsky, S. and Hjelsvold, R., 2002. Visualizing electronic mail. In Proceedings of. the Sixth International Conference on Information Visualisation, 2002. pp. 3–9.

Swartz, N., 2006. New Rules for E-Discovery. Information Management Journal, 40(6), pp. 22–26.

Symantec Corporation, 2011. Enterprise Vault Product Page. Available at: http://www.symantec.com/business/enterprise-vault [Accessed July 4, 2011].

The Sunlight Foundation, 2011. Sarah's Inbox | A Project of the Sunlight Foundation. Available at: http://sarahsinbox.com/ [Accessed July 11, 2011].

Tobias, D.R., 2011. Dan's Mail Format Site. Available at: http://mailformat.dan.info/ [Accessed August 31, 2011].

UK Web Archiving Consortium, 2011. UK Web Archive. Available at: http://www.webarchive.org.uk/ukwa/ [Accessed September 19, 2011].

Universities of Oxford and Manchester, 2008. Paradigm Project Homepage. Available at: http://www.paradigm.ac.uk/ [Accessed September 19, 2011].

University of North Carolina School of Information and Library Science, 2011. LifeTime Library Thought to Be The First Nationwide. *New Release*. Available at: http://sils.unc.edu/news/2011/SILS-lifetime-library [Accessed September 19, 2011].

University of Virginia, 2011. Project Blacklight. Available at: http://projectblacklight.org/ [Accessed August 30, 2011].

US Supreme Court, 2010. *Federal Rules of Civil Procedure*. Legal Information Institute, Cornell University. Available at: http://www.law.cornell.edu/rules/frcp/ [Accessed July 22, 2011].

Vanhoutte, E. and den Branden, R.V., 2009. Describing, Transcribing, Encoding, and Editing Modern Correspondence Material: A Textbased Approach. Literary and Linguistic Computing, 24(1), pp. 77–98.

Viegas, F.B., Boyd, D., Nguyen, D.H., Potter, J. and Donath, J., 2004. Digital Artifacts for Remembering and Storytelling: Posthistory and Social Network Fragments. In Proceedings of the 37th Annual Hawaii International Conference on System Sciences, 2004.. 10 pages. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1265287 [Accessed June 10, 2011].

Villanova University, 2011. VuFind. Available at: http://vufind.org/ [Accessed August 30, 2011].

Van Vleck, T., 2010. The History of Electronic Mail. Available at: http://www.multicians.org/thvv/mail-history.html [Accessed July 6, 2011].

Wagner, F., Krebs, K., Mega, C., Mitschang, B. and Ritter, N., 2008a. Towards the Design of a Scalable Email Archiving and Discovery Solution. In P. Atzeni, A. Caplinskas, & H. Jaakkola, eds. Advances in Databases and Information Systems. Lecture Notes in

Computer Science. Springer Berlin / Heidelberg, pp. 305–320. Available at: http://dx.doi.org/10.1007/978-3-540-85713-6_22 [Accessed June 20, 2011].

Wagner, F., Krebs, K., Mega, C., Mitschang, B. and Ritter, N., 2008b. Email Archiving and Discovery as a Service. In C. Badica, G. Mangioni, V. Carchiolo, & D. Burdescu, eds. Intelligent Distributed Computing, Systems and Applications. Studies in Computational Intelligence. Springer Berlin / Heidelberg, pp. 197–206. Available at: http://dx.doi.org/10.1007/978-3-540-85257-5_20 [Accessed June 20, 2011].

Whittaker, S. and Sidner, C., 1996. Email Overload: Exploring Personal Information Management of Email. In Proceedings of the SIGCHI Conference on Human Factors In Computing Systems: Common Ground. CHI'96. New York, NY, USA: ACM, pp. 276–283. Available at: http://doi.acm.org/10.1145/238386.238530 [Accessed August 10, 2011].

Whittaker, S., Bellotti, V. and Gwizdka, J., 2006. Email in personal information management. Communications of the ACM, 49(1), p. 68.

Wikipedia, Email. Available at: http://en.wikipedia.org/wiki/Email [Accessed August 31, 2011].

Yeh, J.-Y. and Harnly, A., 2006. Email Thread Reassembly Using Similarity Matching. In CEAS 2006 – Third Conference on Email and Anti-Spam. Mountain View, CA. Available at: http://www.ceas.cc/2006/7.pdf [Accessed August 8, 2011].

Yoshinaka, R., 2007. Facing the Changes in the Federal Rules of Civil Procedure | Legal > Civil Procedure from AllBusiness.com. KMWorld, 16(2), p. S7.

Zickuhr, K., 2010. Generations 2010 | Pew Research Center's Internet & American Life Project, Pew Internet and American Life Project. Available at: http://www.pewinternet.org/Reports/2010/Generations-2010.aspx [Accessed June 13, 2011].