



ISSN 1747-1524

DCC | Digital Curation Manual

Instalment on

*“Curating E-Mails:
A life-cycle approach to the management and
preservation of e-mail messages”*

<http://www.dcc.ac.uk/resource/curation-manual/chapters/curating-e-mails/>

Maureen Pennock

UKOLN

University of Bath, BA2 7AY

<http://www.ukoln.ac.uk>

July 2006

Version 1.0

Legal Notices



The Digital Curation Manual is licensed under a Creative Commons Attribution - Non-Commercial - Share-Alike 2.0 License.

© in the collective work - Digital Curation Centre (which in the context of these notices shall mean one or more of the University of Edinburgh, the University of Glasgow, the University of Bath, the Council for the Central Laboratory of the Research Councils and the staff and agents of these parties involved in the work of the Digital Curation Centre), 2005.

© in the individual instalments – the author of the instalment or their employer where relevant (as indicated in catalogue entry below).

The Digital Curation Centre confirms that the owners of copyright in the individual instalments have given permission for their work to be licensed under the Creative Commons license.

Catalogue Entry

Title	DCC Digital Curation Manual Instalment on Curating E-Mails: A life-cycle approach to the management and preservation of e-mail messages
Creator	Maureen Pennock (author)
Subject	Information Technology; Science; Technology--Philosophy; Computer Science; Digital Preservation; Digital Records; Science and the Humanities.
Description	This instalment of the Digital Curation Manual reports on the several issues involved in managing and curating e-mail messages for both current and future use. Although there is no 'one-size-fits-all' solution, this instalment outlines a generic framework for e-mail curation and preservation, provides a summary of current approaches, and addresses the technical, organisational and cultural challenges to successful e-mail management and longer-term curation.
Publisher	HATII, University of Glasgow; University of Edinburgh; UKOLN, University of Bath; Council for the Central Laboratory of the Research Councils.
Contributor	Seamus Ross (editor)
Contributor	Michael Day (editor)
Date	24 July 2006 (creation)
Type	Text
Format	Adobe Portable Document Format v.1.3
Resource Identifier	ISSN 1747-1524
Language	English
Rights	© HATII, University of Glasgow

Citation Guidelines

Maureen Pennock, (July 2006), "*Curating E-Mails: A life-cycle approach to the management and preservation of e-mail messages*", *DCC Digital Curation Manual*, S.Ross, M.Day (eds), Retrieved <date>, from

<http://www.dcc.ac.uk/resource/curation-manual/chapters/curating-e-mails>

About the DCC

The JISC-funded Digital Curation Centre (DCC) provides a focus on research into digital curation expertise and best practice for the storage, management and preservation of digital information to enable its use and re-use over time. The project represents a collaboration between the University of Edinburgh, the University of Glasgow through HATII, UKOLN at the University of Bath, and the Council of the Central Laboratory of the Research Councils (CCLRC). The DCC relies heavily on active participation and feedback from all stakeholder communities. For more information, please visit www.dcc.ac.uk. The DCC is not itself a data repository, nor does it attempt to impose policies and practices of one branch of scholarship upon another. Rather, based on insight from a vibrant research programme that addresses wider issues of data curation and long-term preservation, it will develop and offer programmes of outreach and practical services to assist those who face digital curation challenges. It also seeks to complement and contribute towards the efforts of related organisations, rather than duplicate services.

DCC - Digital Curation Manual

Editors

Seamus Ross
Director, HATII, University of Glasgow (UK)

Michael Day
Research Officer, UKOLN, University of Bath (UK)

Peer Review Board

Neil Beagrie, JISC/British Library
Partnership Manager (UK)

Georg Buechler, Digital Preservation
Specialist, Coordination Agency for
the Long-term Preservation of
Digital Files (Switzerland)

Filip Boudrez, Researcher DAVID,
City Archives of Antwerp (Belgium)

Andrew Charlesworth, Senior
Research Fellow in IT and Law,
University of Bristol (UK)

Robin L. Dale, Program Manager,
RLG Member Programs and
Initiatives, Research Libraries Group
(USA)

Wendy Duff, Associate Professor,
Faculty of Information Studies,
University of Toronto (Canada)

Peter Dukes, Strategy and Liaison
Manager, Infections & Immunity
Section, Research Management
Group, Medical Research Council
(UK)

Terry Eastwood, Professor, School
of Library, Archival and Information
Studies, University of British
Columbia (Canada)

Julie Esanu, Program Officer, U.S.
National Committee for CODATA,
National Academy of Sciences
(USA)

Paul Fiander, Head of BBC
Information and Archives, BBC
(UK)

Luigi Fusco, Senior Advisor for
Earth Observation Department,
European Space Agency (Italy)

Hans Hofman, Director, Erpanet;
Senior Advisor, Nationaal Archief
van Nederland (Netherlands)

Max Kaiser, Coordinator of
Research and Development,
Austrian National Library (Austria)

Carl Lagoze, Senior Research
Associate, Cornell University (USA)

Nancy McGovern, Associate
Director, IRIS Research Department,
Cornell University (USA)

Reagan Moore, Associate Director,
Data-Intensive Computing, San
Diego Supercomputer Center (USA)

Alan Murdock, Head of Records
Management Centre, European
Investment Bank (Luxembourg)

Julian Richards, Director,
Archaeology Data Service,
University of York (UK)

Donald Sawyer, Interim Head,
National Space Science Data Center,
NASA/GSFC (USA)

Jean-Pierre Teil, Head of Constance
Program, Archives nationales de
France (France)

Mark Thorley, NERC Data
Management Coordinator, Natural
Environment Research Council (UK)

Helen Tibbo, Professor, School of
Information and Library Science,
University of North Carolina (USA)

Malcolm Todd, Head of Standards,
Digital Records Management, The
National Archives (UK)

Preface

The Digital Curation Centre (DCC) develops and shares expertise in digital curation and makes accessible best practices in the creation, management, and preservation of digital information to enable its use and re-use over time. Among its key objectives is the development and maintenance of a world-class digital curation manual. The *DCC Digital Curation Manual* is a community-driven resource—from the selection of topics for inclusion through to peer review. The Manual is accessible from the DCC web site (<http://www.dcc.ac.uk/resource/curation-manual>).

Each of the sections of the *DCC Digital Curation Manual* has been designed for use in conjunction with *DCC Briefing Papers*. The briefing papers offer a high-level introduction to a specific topic; they are intended for use by senior managers. The *DCC Digital Curation Manual* instalments provide detailed and practical information aimed at digital curation practitioners. They are designed to assist data creators, curators and re-users to better understand and address the challenges they face and to fulfil the roles they play in creating, managing, and preserving digital information over time. Each instalment will place the topic on which it is focused in the context of digital curation by providing an introduction to the subject, case studies, and guidelines for best practice(s). A full list of areas that the curation manual aims to cover can be found at the DCC web site (<http://www.dcc.ac.uk/resource/curation-manual/chapters>). To ensure that this manual reflects new developments, discoveries, and emerging practices authors will have a chance to update their contributions annually. Initially, we anticipate that the manual will be composed of forty instalments, but as new topics emerge and older topics require more detailed coverage more might be added to the work.

To ensure that the Manual is of the highest quality, the DCC has assembled a peer review panel including a wide range of international experts in the field of digital curation to review each of its instalments and to identify newer areas that should be covered. The current membership of the Peer Review Panel is provided at the beginning of this document.

The DCC actively seeks suggestions for new topics and suggestions or feedback on completed Curation Manual instalments. Both may be sent to the editors of the *DCC Digital Curation Manual* at curation.manual@dcc.ac.uk.

Seamus Ross & Michael Day.
18 April 2005

Biography

Maureen is a Research Officer for the DCC and is based at UKOLN at the University of Bath. She has been involved in digital preservation research since 2001 when she joined the Dutch government *Digital Preservation Testbed* research project. Maureen also worked at the Dutch National Archives on the EC-funded ERPANET project, and on the e-Government Knowledge Centre project at the Dutch ICT Foundation (Stichting ICTU). Her main areas of interest lie in strategies for digital preservation, digital record-keeping and digital archiving, digital cultural heritage, and authenticity of digital records. She joined the DCC in Autumn 2005.

Table of Contents

1. Introduction and scope	7
2. Curating E-mails	8
2.1 What is digital curation?.....	8
2.2 Why do e-mails need curating?.....	8
2.3 Authenticity and Integrity of e-mails.....	9
2.4 E-mail as a record.....	10
2.5 E-mail as a valuable cultural, historical, or research object.....	14
3. E-mail Curation: Roles and Responsibilities	15
3.1 Issues for Creators and Recipients of e-mail messages.....	15
3.2 Issues for Curators of e-mail messages.....	16
3.3 Issues for Re-users of e-mail messages.....	21
4. Preserving E-mails	24
4.1 Options for preserving e-mail messages.....	24
4.2 Dealing with Attachments.....	27
4.3 Dealing with Digital Signatures.....	28
4.4 Preservation Metadata.....	28
4.5 Long term storage and archiving.....	29
5. Practical Steps	31
5.1 Developing an e-mail policy.....	31
5.2 Educating users and stakeholders on their responsibilities in e-mail curation.....	32
5.3 Implementing a solution to capture and preserve e-mails for specific retention periods.....	33
6. Future Developments	34
7. Conclusion	37
Glossary	38
References	39
Key external resources	44
Appendix 1 – Sample institutional guidelines, advice, and policies on managing e-mails	45
Appendix 2 - E-mail Curation and Preservation In Action	48
A2.1 Technological solutions for e-mail preservation.....	48
A2.2 E-mail re-use in publicly available collections.....	52

1. Introduction and scope

The rise and proliferation of digital technologies has resulted in an expansion of opportunities for institutions to create, manage, and maintain records and documents in digital form. One of the forms these records often take is e-mail. The first electronic message – e-mail for short – was sent in the 1960's over a single mainframe system and network e-mail followed shortly after in the early 1970's. Sadly, the message that was sent was not recorded for posterity. The subject of the message is unclear and the contents of the message have been lost to the so-called 'digital dark ages'.

This initial failure to preserve was indicative of what was to follow: modern companies in the twenty-first century have collectively been fined billions of dollars for failing to adequately manage and preserve corporate e-mail records, and others have been similarly fined for creating and retaining inappropriate e-mail records.¹ Despite the fact that e-mail forms the backbone of communications in many modern institutions and organisations, it

is often badly managed and the long-term preservation of e-mail messages is a challenge for which most solutions have yet to be put to the test. Beginning with creation, and continuing through to long-term preservation or disposal of messages, e-mail curation in most organisations faces the same problems as it did when the medium was first developed. E-mail is thus both a solution and a challenge in the modern working environment.

This instalment of the Digital Curation Manual will report on the several issues involved in managing and curating e-mail messages. Although it is not possible to offer an immediate 'one-size-fits-all' solution, the instalment will outline a framework for e-mail curation and preservation, provide a summary of current approaches, and address the organisational and human challenges to successful e-mail curation. The instalment also offers some thoughts on the future development or evolution of e-mail. This is inevitable, given the rapid pace of technological advancement over the past fifty years and the current changes in messaging technologies.

¹ In May 2005, Investment Bank Morgan Stanley were ordered to pay \$1.45 billion in damages in a case referred to by some as a 'legal chernobyl', after failing to make e-mails available in a legal case, see <http://www.silicon.com/research/specialreports/compliance/0,3800003180,39130615,00.htm>. In a related case, they have recently offered the U.S. Securities and Exchange Commission (SEC) \$15 million in an attempt to settle an investigation arising from this legal dispute, see <http://www.out-law.com/page-6656>. In December 2002, the SEC, the New York Stock Exchange and NASD fined five U.S. Companies - Deutsche Bank Securities Inc.; Goldman, Sachs & Co.; Morgan Stanley & Co. Incorporated; Salomon Smith Barney Inc.; and U.S. Bancorp Piper Jaffray Inc a total of \$8.25 million for failure to preserve e-mail communications, see <http://www.sec.gov/news/press/2002-173.htm>. Merrill Lynch also faced fines of over \$100 million in 2002 after keeping inappropriate e-mails that it was subsequently forced to disclose, see <http://www.internetnews.com/bus-news/article.php/1551141>.

2. Curating E-mails

2.1 What is digital curation?

The realisation that preservation and archiving are but stages in a larger chain of events has led to adoption of the phrase 'digital curation' to describe the management and preservation of digital data. Digital curation is not simply a matter for those charged with care of resources at the end of their active lives, for the term 'digital curation' refers to the ongoing management of digital materials for both current AND future use. Curation issues are relevant from day one of the records life-cycle, from creation through to curation and including re-use of the data.

“Digital Curation: The activity of, managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purpose, and available for discovery and re-use. For dynamic datasets this may mean continuous enrichment or updating to keep it fit for purpose. Higher levels of curation will also involve maintaining links with annotation and with other published materials.”

- Lord & MacDonald (2003)

Digital Curation therefore requires action from several different stakeholder groups, such as the data creators or originators, record-keepers, archivists, librarians, and IT specialists (these four are hereafter referred to collectively as data curators unless otherwise noted), and data re-users. Organisations that should carry out e-mail curation include not only archives and libraries, but also record-creating bodies such as government offices (i.e. e-government), businesses (i.e. e-commerce), HE/FE institutions, and projects where e-mail constitutes a valuable record of project developments. Individuals with valuable e-mail

collections should also carry out e-mail curation activities to ensure their records, which may have cultural, heritage, or scientific value, can be made available to future generations.

Unfortunately, the organisational curation of e-mail messages is often overlooked as a) it is a new type of record that is often not properly integrated into an overall record-keeping infrastructure, and b) responsibility for e-mail curation is not explicitly allocated to specific staff and is only addressed implicitly. This is contrary to websites, where web-masters have a designated responsibility to manage the Internet content and where international activities and awareness of curation needs are higher.

2.2 Why do e-mails need curating?

All e-mails must be properly curated over their life-span if they are to remain usable and authentic for future (as yet unknown) re-users. Like other digital records, e-mails rely on a combination of hardware, software, and content files in order to be rendered on-screen. Despite the use of a single standard – RFC 2822 – that forms the basis of successful transmission of e-mail messages between different e-mail clients, this standard is not widely acknowledged outside technical circles.² Many e-mail clients convert the standard file into a proprietary or non-standard format for storage, which can make it difficult to access the messages if the client becomes unavailable. Messages are also often badly created, with insufficient contextual information to identify the context and authors of the message in the future, and badly managed – transient messages are kept

² FRC 2822 Internet Message Format - <http://www.faqs.org/rfcs/rfc2822.html>.

alongside messages with valuable contents, filing and categorisation of e-mails is inconsistent between and within users, and 'sent' mails are often completely ignored. These issues can all affect the viability and stability of the messages, not only for future users over the long-term, but even in current and active use.

Many organisations depend upon their disaster recovery strategy and backup tapes for access to old e-mails and e-mail archiving. However, backup tapes are fundamentally different to a proper electronic archive. Backups have no logical file structure and simply consist of data physically streamed to a series of one or more tapes. Unofficial and personal messages are mixed in with official organisational records, varying retention periods are not applied, there is no additional metadata to explain message context, relationships, or facilitate retrieval, and the long-term preservation of the messages and their attachments is not addressed. Backups, quite simply, are not a suitable solution for curating, archiving or preserving e-mails.

Good curation practices offer a framework in which the viability and reliability of the messages can be secured into the future.

2.3 Authenticity and Integrity of e-mails

Digital objects are easily mutated. Their value rests on the proven authenticity and integrity of the files and their rendered content; as such, curators must be able to prove that an object has not been tampered with. It is easier to tell when tampering has taken place on traditional paper-based objects because the 'form' of the object is permanently fixed on a tangible medium and to convincingly tamper with or forge such objects requires significant skill and dedication. Altering intangible digital objects does not require the same level or type of skill, so special steps must be taken to ensure that e-

mails are created, stored, and maintained in a manner that can help ensure their authenticity and integrity through time.

Authenticity and Integrity are two sides of the same coin. An object is authentic when proven that it is what it purports to be. Creating and retaining metadata about the object's origins and chain of custody can help prove this. Integrity pertains to the contents of the object: objects are integrally whole when the information that they contain is complete and unaltered in all essential aspects. Both authenticity and integrity must be preserved for the value of the object to be maintained.

Several steps can be taken to ensure that the authenticity and integrity of e-mails can be preserved. E-mails can be created with organisational contextual metadata in their content that will help prove their authenticity. Similarly, preserving the RFC 2822 file in its entirety will ensure metadata that charts the transmission path of the e-mail is preserved - this metadata establishes that the message was sent from the purported sender to the purported recipient and identifies the date and times it passed through each e-mail server, thus also addressing authenticity. Most e-mail servers keep a log-file of e-mail traffic, which if kept can be used to confirm which messages were sent, by whom, and when. Implementing security to limit access to the stored messages and maintaining an audit trail of actions by restricted and identified individuals helps ensure that the integrity of the messages is not purposefully compromised. Finally, actively monitoring the preserved messages over time to avoid bit deterioration and technological obsolescence ensures that the integrity of the messages is not accidentally compromised.

Authenticity and Integrity are central concepts in archival science and records management. For long-term preservation of reliable e-mail messages, it is essential that they are also

recognised in the complimentary fields of librarianship and IT.

2.4 E-mail as a record

When e-mail was first introduced to institutions, many saw it as an informal or personal means of communication. Its distributed nature lent itself to this perception. E-mail messages were frequently not considered relevant in records scheduling, and there was little guidance on using e-mail systems, let alone on preserving messages. However, recent legal cases concerning inappropriate e-mails and e-mail mismanagement have resulted in fines totalling billions of dollars for the companies concerned. As a result, the status of e-mail as a formal records mechanism is now more widely recognised.

The basic definition of a record is anything that documents a working transaction between two or more parties, that documents the mission and goals of an organisation, or that was created or received in the course of carrying out the mission and goals of an organisation. The BS ISO 15489 Standard defines a record simply as: “any information that is created, received and maintained as evidence and information by an organisation or person in pursuance of legal obligations or in the transaction of business”.³ E-mail records formed a vital part of the evidence supplied to the UK Hutton Inquiry in 2003/2004, which investigated the circumstances surrounding the death of civil servant Dr David Kelly after he spoke to a journalist about the British Government's support for commencement of the war in Iraq.⁴ Most records managers will

now attest to the fact that e-mails can be records, without question.

One of the primary stumbling blocks in successful institutional e-mail curation arises from the distributed nature of the medium: messages are sent and delivered to users desktops without any obvious interference by a hosting institution. This can lead to the (incorrect) perception that e-mail messages are a personal matter and not one for the institution. Records managers must give guidance to record originators on identifying electronic records stored on their PC's or in their personal file space. This is particularly necessary for identifying e-mail messages that constitute records, as many of the messages in a users inbox are ephemeral, transient, or copies of records for which the responsibility to preserve lies elsewhere and such messages can be discarded. A simple but useful initial decision model to identify e-mail records from non-records can be found in advice from the Dutch National Archives.⁵

For many institutions, record keeping is a legal obligation and it is this that drives their record-keeping activities. Many Acts of the UK and Scottish Parliaments require that organisations keep records, including e-mails, and that they observe certain other legal obligations and statutes relating to those records.⁶ UK

inquiry.org.uk/index.htm contains a wealth of information about this incident. A list of evidence supplied to the Inquiry is available at <http://www.the-hutton-inquiry.org.uk/content/evidence.htm>. The e-mails available on the website are electronic versions of printed e-mails but there is no information on the website to indicate how their authenticity and integrity was verified.

⁵ See 'From digital transience to digital durability; Preserving E-mails' from the Digitale Duurzaamheid project of the Dutch National Archives.

<http://www.digitaleduurzaamheid.nl/home.cfm>.

⁶ An excellent overview of legal issues pertaining to e-mail archiving from a European perspective can be found in the paper from the Antwerp City Archives: *Archiving E-mails*, (2002) by Boudrez F & Eynde, Sofie

³ BS ISO 15489-1:2001 Standard for Information and documentation: Records management. Available from <http://www.bsi-global.com/ICT/Legal/bsiso15489-1.xalter>

⁴ The Hutton Inquiry website <http://www.the-hutton->

legislation has been taken as the basis for the following section, which discusses how specific UK and European Acts relate to the management curation of e-mail messages.

Data Protection Act (1998)

The 1998 Data Protection Act sets out eight principles that govern the use of personal information.⁷ All organisations must comply with these principles, which apply to e-mail as equally as to other record types. Particularly relevant for e-mail messages are principles 4, 5, and 7:

4. Personal data shall be accurate and, where necessary, kept up to date.
5. Personal data processed for any purpose or purposes shall not be kept for longer than is necessary for that purpose or those purposes.
7. Appropriate technical and organisational measures shall be taken against unauthorised or unlawful processing of personal data and against accidental loss or destruction of, or damage to, personal data.

Rules on retention (principle 5) preclude organisations from simply carrying out a mass harvest of e-mail messages and storing them for an arbitrary period of time. Requests must be answered within forty calendar days of receipt, so good management and curation of e-mails is necessary to enable relevant data to be located within that period.

Freedom of Information Act (2000)

The Freedom of Information (FoI) Act (2000) and the Freedom of Information (Scotland) Act

Van den. Available from <http://www.expertisecentrumdavid.be/davidproject/teksten/Rapporten/Report4.pdf>.

⁷ The full text of the Data Protection Act can be obtained from the Office of Public Sector Information (OPSI) website:

<http://www.opsi.gov.uk/acts/acts1998/19980029.htm>.

(2002) gives people the right to access data held by a public authority that is not already covered by the Data Protection Act, unless the data is exempt from disclosure.⁸ Organisations must respond to requests received under the FoI Act within twenty working days. E-mail is included as relevant data under this Act but e-mail discovery carries some inherent problems: data relevant to a FoI request may or may not be indicated in a subject line, it may be in the message content alone, or simply in an attachment/embedded file of any format; e-mails simply stored in a user's Inbox are not indexed for fast searching, and; mails can be stored in numerous locations, not all of which are accessible to systems administration or record keepers. Given the phenomenal number of e-mails an institution can accumulate in the course of day-to-day operations, and the time limit of twenty days for organisations to respond to requests, it is vital that sufficient records management practices are in place so as to be able to fully comply with the request within the given term without incurring huge re-discovery costs and a significant increase in workload.

Good record keeping practices will inevitably involve regular deletion of e-mails that do not require retention, however, e-mails must not be deleted with the intention of evading disclosure. Moreover, e-mails that must by law be deleted must be completely destroyed. Simply 'deleting' from a user's inbox or the institutions live system is insufficient: back-up copies or messages in the recycle bin may later be discovered and thus still form part of an organisation's electronic information.⁹ To fully

⁸ The full text of the Freedom of Information Act is available from the OPSI website at <http://www.opsi.gov.uk/acts/acts2000/20000036.htm>. The Freedom of Information (Scotland) Act is available from the Scottish Information Commissioner's website: <http://www.itspublicknowledge.info/legislation/act/foiact/contents.htm>.

⁹ Communications from the Information Commissioner, the Department of Constitutional Affairs, and the

comply with the principles of both the Freedom of Information Act and the Data Protection Act, a data cleansing (or destruction) strategy that will ensure inappropriate or unnecessary data is properly destroyed, is just as essential as a data retention strategy.

Regulation of Investigatory Powers Act (2000)

The Regulation of Investigatory Powers (RIP) Act defines the extent to which an organisation can monitor and access communications, including e-mails, created or received by personnel in the course of their work.¹⁰ This applies regardless of whether the employer is a public or private body. Incorporating e-mail curation into a wider organisational records management infrastructure will often involve access to messages by persons other than those named on the message headers. In such instances, all parties must be made aware of the extent to which e-mail messages are monitored and archived, especially if archiving is automated and not initiated by the message creator/recipient.

Human Rights Act (1998)

The Human Rights Act came into force in 2000.¹¹ It incorporates the European Convention on Human Rights into UK law, making civil and political rights enforceable by courts in England and Wales.¹² The

Freedom of Information Tribunal are inconsistent, but the current stance appears to be that data in 'deleted' files such as the recycle bin or on backup tape may still be considered to fall under the confines of the FoI Act. See article in The Register, Dec 2005, http://www.theregister.co.uk/2005/12/22/foia_undelete_ruling/.

¹⁰ The Regulation of Investigatory Powers Act is available from the OPSI website at <http://www.opsi.gov.uk/acts/acts2000/20000023.htm>.

¹¹ The UK Human Rights Act is available from the Office of Public Sector Information <http://www.opsi.gov.uk/acts/acts1998/19980042.htm>.

¹² The European Convention on Human Rights is available from the website of the European Court of

Convention stipulates that everyone has the right to respect for his private and family life, his home and his correspondence. It also states that there shall be no interference by a public authority with the exercise of this right unless in accordance with the law and specifies circumstances in which such exceptions could be legitimate.

It has been argued that under the terms of this convention all e-mail is confidential in principle. Archiving of employee e-mail by an employer or records manager is therefore an invasion of privacy.¹³ Whilst the UK RIP Act makes clear-cut provisions for interception of e-mails for the purpose of monitoring or recording business communications, national law is always superseded by European Law. Employers therefore need to clearly establish and communicate the extent to which e-mails will be monitored and recorded, so that employees have no expectation of privacy in this regard.

Other legal issues

E-mails and collections of e-mails can also be affected by other legislation, for example, intellectual property legislation. Three main aspects of intellectual property protection are discussed here: trade secrets, copyright, and database right.

'Trade Secrets' are a valuable form of intellectual property and protection for Trade Secrets arises from the Law of Confidentiality.¹⁴ The Law of Confidentiality provides remedy for the unauthorised use or disclosure of information that is confidential

Human Rights <http://www.echr.coe.int/echr>.

¹³ This point was argued in Boudrez et al, 'Archiving Emails', op cit. A similar concern has been made in a confidential report from the UK investigating automatic archiving of all e-mail in and out of an institution.

¹⁴ See http://www.intellectual-property.gov.uk/resources/other_ip_rights/trade_secrets.htm for more details.

and prevents illegitimate use of confidential information by a recipient of that information. There may be an obligation of confidentiality, for instance, between a lawyer and a client, between employers and employees, or between healthcare professionals and clients. When confidential information is communicated by e-mail, the law of confidentiality must still be observed. The Law of Confidentiality is largely a part of Common Law and whilst common law is not written as an Act of Parliament, it is nonetheless a major part of the law.

Closely related to this is Copyright Law.¹⁵ Although the subject of the intellectual ownership of the contents of an e-mail (and its attachments) have not yet been explored in great detail, copyright issues may prevent the contents of an e-mail from being publicly circulated. This is a significant factor in re-use or re-distribution of archived or stored e-mail messages and will no doubt depend on the context of the message and the re-use environment.¹⁶

Database Right is an intellectual property right that can be applied to a collection of e-mails. A database is defined as 'a collection of data or other material that is arranged in such a way so that the items are individually accessible'.¹⁷ By this definition, a collection of e-mail messages can be classified as a database and therefore

¹⁵ See

<http://www.patent.gov.uk/copy/legislation/copylaw.htm> for copyright-related legislation. This site is hosted by the UK Patent Office. An unofficial, consolidated text of UK legislation relating to copyright, performance rights, publication right, and database right is also available from the UK Patent Office at <http://www.patent.gov.uk/copy/legislation/legislation.pdf>

¹⁶ The JISC-funded Paradigm Project is exploring some of these issues as part of its research into the issues involved in preserving digital private papers. More information is available from the project website, <http://www.paradigm.ac.uk/>.

¹⁷ See http://www.intellectual-property.gov.uk/resources/other_ip_rights/database_right.htm for more details.

afforded the same protection. Database Right is very similar to Copyright in that it is automatic, but the term of protection is much shorter and the activities that a rights holder can control are different.

Risk Management

The main risks of failing to properly manage and curate e-mail messages include:

- Legal consequences
- Financial consequences
- Loss of public credibility
- Loss of organisational memory
- Loss of accountability
- Lack of transparency (particularly important for government organisations)
- Failure to exploit existing digital resources
- Failure to provide useful services to users

The specific risks of failing to manage organisational e-mail messages will differ slightly according to the type of organisation and its remit. Many of the risks above can be mitigated by the implementation of a risk management approach, which is fundamentally based on good records management principles.

A risk management framework is best developed as a result of an information compliance audit and risk assessment exercise. This will enable an organisation to determine the likelihood of threats arising from their e-mail management approach and the impact of such threats should they be realised. It will also identify what level of risk is acceptable and may indicate which tools, techniques, and processes are most appropriate to address the risks.

2.5 E-mail as a valuable cultural, historical, or research object

Correspondence of notable figures is traditionally collected and used by institutions, private individuals, and researchers as a historical record of events or social/private discourse. There is often no legal reason why the correspondence is created or should be preserved. The value and preservation of the materials arises instead from its cultural heritage and historical context.

E-mail is the modern-day equivalent of paper-based correspondence. Many types of institutions, particularly in the library and museum sectors, will therefore collect and curate e-mails that have no legal context or value, and with no explicit legal obligation to do so. Such organisations may simply have an implicit or explicit mandate to preserve. For example, a library may accession a collection of e-mails from a significant author that it wishes to curate and make accessible for future

generations. Record-keeping is not the driver for managing and curating objects in such institutions; the driver instead is the preservation and accessibility of the objects themselves. In the absence of legal requirements, the institution's mandate to provide continuing storage and access to items is itself a powerful impetus for ensuring proper e-mail curation practices.

In other situations, e-mail collections from collaborative and distributed research projects may be retained, such as those in the field of e-Science, HE, or collaborative EU projects. Such collections are valuable not so much for their cultural heritage value, but to demonstrate how results were disseminated and the process of analysis that led to the project conclusions. They often explicitly chart development of the project and can be used to demonstrate to funders the attempts that were made to meet project goals and deliverables.

3. E-mail Curation: Roles and Responsibilities

3.1 Issues for Creators and Recipients of e-mail messages

Curation and preservation begins at source. Creators and recipients of e-mail messages are therefore the first in a chain of important users. Notwithstanding the legal issues that e-mail creators and recipients in an institutional or business context must be aware of, there are several other important issues for users to consider so that e-mails are created and managed in a manner to enable their authenticity and integrity to be proven through time.

Creating e-mails

Institutional e-mail creation practices are often left to the individual to decide – selecting plain text or html, using a signature or no signature, to CC or BCC, and to attach a file or insert a link – and messages are therefore often created in an ad-hoc manner. An absence of institutional guidance on creation practices perpetuates the perception that e-mail is a personal issue. Institutional guidance and good practice guides not only convey the message that e-mail is a business tool, they also help ensure the creation of stylistically consistent messages. This in turn affects the level of effort required to successfully preserve messages, in that preserving large batches of messages will be more straightforward if they all conform to an expected pattern.

There is a critical point at which users decide how they will use the e-mail application for communicating or transmitting specific types of content. This decision can impact not only on the number of e-mails that must be retained, but also on the content of the e-mails. In many cases, 'official' and documentary content is not contained in the body of the e-mail message but is affixed to the message in the form of an

attachment. For example, letters of reference and policy documents may be transmitted by e-mail but they are actually text documents affixed to the e-mail; in these instances, the e-mail is used as a carrier and the actual documentary records (the letter or policy document) are stored, managed, and 'archived' through a different mechanism. Such instances beg the question of whether or not the e-mail, which in such scenarios can be likened to an envelope, ought to be preserved, or whether it suffices to preserve only the attached file. This is a issue on which curators must provide guidance for creators, not only regarding creation of the e-mail message, but also on submission of the e-mail message for preservation.

Metadata for e-mails

E-mail transmission files include a significant amount of header metadata – more than is usually displayed by the e-mail client - that can be used to prove the authenticity of the message and its provenance. Headers such as 'To', 'From', 'Received by', and 'Date' are all useful in determining the transmission path of an e-mail and the different files that may have been attached. The oft-cited U.S. Department of Defense Standard 5015.2 “Design Criteria Standard For Electronic Records Management Software Applications” identifies a basic set of essential header elements that should be retained:

- The intelligent name of the sender.
- The intelligent name of all primary addressees (or distribution lists).
- The intelligent name of all other addressees (or distribution lists).
- The date and time the message was sent.
- For messages received, the date and time the message was received (if

available).

- The subject of the message.¹⁸

This is largely consistent with Records Management Application requirements from the UK National Archives, although the terminology differs.¹⁹ However, as mentioned, the e-mail transmission file will usually contain additional header fields and organisations should ideally attempt to retain as much of this transmission data as possible. This data can then be reinforced or supplemented by the data held in an e-mail server traffic log, if necessary.

Contextual metadata in the message header can be enhanced by using the address book linked to an e-mail application, which enables the full name (or 'intelligent name' in DoD standard 5015.2 terminology) of recipients to be appended to the e-mail address. This can help prove the identity of recipients at a later date if the e-mail address alone is not enough to identify them. Institutional guidance may be required to achieve this and is possible without customisation of the e-mail client. Further contextual and administrative metadata can be added to the message headers by customising the e-mail client. If these additional headers are made compulsory and content controlled by the

use of drop down lists, they can also be used to filter messages to the correct storage folders. All of this header metadata is a valuable source for automatic metadata extraction.

Message creators should be encouraged to include additional metadata in the message content, if possible. Using a signature block in the message content is a simple but effective way to achieve this and add contextual metadata.

Received messages & Inbox management

Due to the sheer volume of e-mails exchanged between users each day, e-mail inboxes can quickly become unmanageable if automatic message sorting and filtering is not implemented. This task can be largely automated, although regular human intervention is invariably required to ensure transient or temporary mails are properly deleted and that an adequate and useful folder system is established.

Compulsory storage of official messages in shared storage locations and institutional use of IMAP rather than POP3 can reduce the risk that official messages in user's inboxes will be overlooked.²⁰

3.2 Issues for Curators of e-mail messages

One of the initial tasks for curators and institutions with a responsibility for curating e-mails is to offer guidance to creators on the creation of e-mail records and the management of email inboxes. However, as curatorial responsibilities permeate across the entire life cycle of the data objects, curating digital

¹⁸ U.S. Department of Defense Standard 5015.2 'Design Criteria Standard For Electronic Records Management Software Applications' (2002)

<http://jitic.fhu.disa.mil/recmgt/p50152s2.pdf>. The *MoREQ (Model Requirements for the Management of Electronic Records) Specification* also recommends the retention of 'intelligent' names, in that it is preferable to retain the name of a correspondent in full rather than simply their e-mail address. The MoREQ Specification was prepared for the IDA Programme of the European Commission in 2001, see <http://cornw.co.uk/moreq.html>.

¹⁹ See particularly the *Requirements for Electronic Records Management Systems*, <http://www.nationalarchives.gov.uk/electronicrecords/reqs2002/pdf/referencefinal.pdf>, part of The National Archives' *Functional Requirements for Electronic Records Management Systems* (2002) at <http://www.nationalarchives.gov.uk/electronicrecords/reqs2002/>.

²⁰ IMAP and POP3 are two alternative message delivery protocols. IMAP is optimised for message access on the server; when using POP3, users connect to the server only briefly and download messages to their individual machines. Most institutions and web mail services use IMAP, whereas personal home accounts often use POP3.

resources involves collaboration with staff across the entire breadth of institutions, not just data creators.

E-mail policies

Policy indicates the extent to which activities are embedded in an organisation and the importance attributed to them. Many organisations have e-mail policies, but these have historically related primarily to the *use* of e-mail and cover terms of allowable use, security, and confidentiality. These policies are increasingly being expanded to incorporate interception conditions and basic archiving and retention procedures, yet they rarely cater for actual preservation of the messages in digital form.

The AIIM 2003 Industry Watch “*E-mail Policies and Practices: An Industry Study Conducted by AIIM International and Kahn Consulting, Inc., 2003*” surveyed over one thousand respondents from a range of industries in the US:

- “70% of organizations tell their employees what to expect in terms of the privacy of e-mail at work,
- 80% dictate acceptable use of the e-mail system
- 73% provide guidelines on e-mail content.
- 60% have NO formal policy governing its retention;
- 54% do not tell employees where, how, or by whom e-mail messages should be retained.
- When organizations do retain e-mail messages, only 37% retain messages according to their content.
- 31% keep e-mail indefinitely
- 26% retain it for less than 120 days.
- 67% use maximum mailbox sizes as a method of creating a de facto retention limitation.”

They found that whilst one hundred percent used e-mail for business purposes, less than eighty percent had a formal written policy regarding the use of e-mail. Their findings confirm that although an increasing number of organisations are developing policies on use and privacy, there remains a great deal of work to be done regarding policies and strategies for retention.

A series of case studies carried out by the ERPANET project from 2002 to 2004 found that most institutions do not have specific policies for preservation of their digital assets. Some institutions are starting to create preservation policies, but it is rare that e-mail has a preservation policy of its own.²¹ To date, the DCC is aware of very few policies specifically relating to the curation of digital materials, perhaps as curation for digital materials is a concept that has only recently emerged.²² However, as digital curation is essential for the creation and maintenance of digital organisational memory and accountability, it would appear prudent to establish such a policy to provide what could otherwise be considered a nominal and supporting activity with the high-level backing often reserved for core business functions. Such a policy may address at the very least use,

²¹ Studies are available from <http://www.erpanet.org/studies/index.php>. A notable exception to this was the National Library of Wales, which had established policies specifically for digital preservation and electronic record keeping and had integrated them into the wider policy framework of the entire institution.

²² However, staff from the Public Record Office of Northern Ireland reported on development of a digital curation policy when attending the DCC Information Day in Belfast on December 1st 2005. A report of this session is available at http://www.dcc.ac.uk/training/info-day-2005-december/belfast_info_day.pdf. Furthermore, curation is emerging as a significant factor in Research Council data sharing and preservation policies: see Marc Thorley’s overview presentation from the DCC/DPC workshop on Policies for Preservation and Curation (July 2006) at <http://www.dcc.ac.uk/events/policy-2006/>.

management and preservation. It is not necessary to establish a curation policy uniquely for e-mail messages as a generic policy on digital curation would probably suffice; however, integrating the policy into the wider policy framework of the organisation is essential for it to be adopted by all employees. In the same manner, backing from senior management will help achieve the policy's aims.

Selecting e-mails for long term curation and preservation

Selection is the process of deciding what will be added to a collection. In a historical or cultural context, where e-mail collections are developed without legal record-keeping requirements, there is increased merit in the 'keep everything' approach. The historical value of different e-mails will vary according to needs of the researcher using the collection, but they will **all** be of some value.²³

Furthermore, retention of the entire set of e-mails enables more complete investigation and analysis of the collection, and the argument that a collection should not be fragmented is particularly relevant here. However, such organisations (for example, libraries) may still carry out selection activities regarding which collections to retain, rather than which aspects of a collection to retain. Selection criteria influencing such decisions are traditionally developed based on five key criteria: evidential value, aesthetic value, market value, associational value, and exhibition value.²⁴

Institutions that do have legal record-keeping

²³ Furthermore, the retention of an entire collection allows analyses (for example of social networks and collaborations) that simply cannot be reliably carried out on a limited selection of messages.

²⁴ From Ross Harvey's forthcoming DCC Manual chapter on Appraisal and Selection, which will be posted at <http://www.dcc.ac.uk/resource/curation-manual/chapters/appraisal-and-selection/>. Readers are referred to this chapter for more detailed information on the topic.

responsibilities should have institutional selection and retention schedules that identify the types of records to be preserved and the preservation term.²⁵ Not all e-mails have to be preserved: most are transient and can be deleted almost immediately and many others have short-term retention value and can be deleted after periods ranging from six months to a few years. Usually, only a small percentage require preservation for a longer period. Users often require guidance to identify official records that require retention against official transient messages. Appraisal and selection of messages should ideally be made at source and at the point of receipt or creation, so that official e-mails are immediately separated from private or transient ones and entered into the organisation's record keeping infrastructure. Early implementation of selection decisions also helps minimise the risk that e-mails will be lost, accidentally deleted, or deteriorate.

In a business context, encouraging people to engage in planned retention is preferable to the 'keep everything' approach. Non-selective archiving of all e-mails increases the costs of storage and research indicates that it also increases the amount of time it takes to locate objects.²⁶ It also leaves organisations susceptible to legal problems arising from, for example, retention of inappropriate messages. E-mails that do not have to be kept, or which must be deleted for compliance with legal requirements, must be properly destroyed. This includes copies of messages stored on back-up

²⁵ A generic policy for e-mail retention and disposal was developed by the JISC Institutional Records Management and e-mail project, see <http://www.lboro.ac.uk/computing/irm/generic-policy.html>.

²⁶ D. Reier, *I have to show them what?! E-mail and the process of electronic discovery*, in *Information Storage and Security Journal*, June 2005, as cited in Boudrez, Filip, 'Filing and Archiving E-mail' (2006) http://www.expertisecentrumdavid.be/docs/filingArchiving_email.pdf.

tapes that are kept for business continuity purposes.

E-mail selection criteria should consider whether entire threads should be preserved and optimum thread structuring possibilities. Decisions on how best to preserve the structure of digital objects are not limited to e-mails. All types of digital objects can have different attributes (or characteristics) to their traditional paper-based counterparts and decisions must be taken on the attributes that must be preserved in order for an object to remain authentic and integrally sound. Not all of the original attributes are necessarily required.

Determining authenticity requirements for e-mail preservation

Authenticity requirements play a crucial role in the selection of a preservation approach. The requirements refer not only to the records concerned, but also to the environment in which the records are stored.

Digital records have at least four and possibly even five attributes. All records have Content, Context, Structure and Appearance. Structure and appearance are often linked together and referred to as a single attribute. Although this was usually true for traditional records, some elements of each are independent of the other when dealing with digital records: for example, structure can be recorded by mark-up that does not rely upon rendering of the record in order for the structure to be understood; likewise, there are aspects of appearance such as colour and formatting that may have nothing to do with structure but are indicative of, for example, emphasis or special denotation. Some records also have Behavioural attributes that allow users to carry out certain actions, such as running a spreadsheet formula to carry out calculations, or invoking an embedded macro. These often rely on the functionality of the application used to render and use the record and are not necessarily stored as part of the

record file.

Not all attributes are necessarily required for successful preservation. Decisions must be taken on the requirements of the institution, its future users, its capabilities, and the record type itself – what attributes are necessary in order for the record, i.e. e-mail, to be understood and used in the future as a faithful rendition of the record when originally created and used? Generic guidelines are available and form a useful starting point for institutions to formulate their own requirements. The International Council on Archives' (ICA) Guide for Managing Electronic Records from an Archival Perspective²⁷ considers digital objects to be comprised of only three attributes - content, context and structure – and requires preservation of all three to ensure authenticity over time. The Digital Preservation Testbed of the Dutch National Archives has carried out research specifically into e-mail preservation which indicated that content, context, structure and appearance should all be preserved.²⁸ The importance of preserving structure in e-mails is twofold: firstly, the structure of the e-mail message, comprised of headers, message body, and attachments, must be preserved; secondly, the structure of a series of e-mails that together form an e-mail thread may also require preservation. Implicit in this is the preservation of 'understanding' - it must be clear who said what to whom. This can be difficult when people reply to a group e-mail and include the original text in their reply but insert comments in several points in the original text rather than inserting their reply as a block of text. Such e-mails are easy to understand when this happens only a few times between two people, but

²⁷ The ICA *Guide for Managing Electronic Records from an Archival Perspective* is available from <http://www.ica.org/biblio.php?pdocid=3>.

²⁸ The Digital Preservation Testbed developed a set of authenticity requirements for preservation of government e-mail records. These are a useful starting point from which other institutions may develop their own, institution-specific, authenticity requirements.

quickly become a complicated mess of unintelligible comments when contributed to by several people and accessed again at a later date. Curators should thus address this when issuing advice on message creation.

Audit trails can be used to prove that records remain authentic and have not been tampered with, so establishing an audit trail may also constitute an authenticity requirement. Audit trails can be established from the point of accession into an archive or collection, and should accompany the record or records until disposal. Audit trails for preservation may take the form of a preservation log book, with metadata showing who accessed a given computer system, the actions that were carried out, when the activities were undertaken, and any change that resulted in the records. (The National Archives of Australia have developed software specifically for this purpose, see section Appendix 2.)

Integrating e-mail preservation into a larger overall records and archives management strategy

One of the main challenges in e-mail preservation arises from the distributed nature of e-mail messages. E-mails are often not integrated into a wider record-keeping system and are thus prone to loss. Even when e-mails are saved to a shared folder in an e-mail system, they are often stored separately from other electronic records and so multiple sites must be accessed and checked to make sure that all relevant records are retrieved when an FoI request is received or when a scheduled migration to an external repository is due. Storage of institutional records on a shared server, in a Records Management System (RMS) with other electronic records, or in a digital repository is the best way to avoid these issues and is the most efficient way to manage a large heterogeneous records collection.

E-mail records are sometimes transferred into a

dedicated e-mail repository and several commercial e-mail storage solutions have emerged in recent years. This may be largely due to the disproportionate amount of press coverage that e-mail has received in the course of regular records disclosure, whereby e-mail is seen as the problem area of records management and therefore as a particularly lucrative market. The sheer volume of e-mails sent and received by institutions is another contributing factor to the emergence of this market.²⁹ Their suitability for ongoing curation however, has yet to be proven. Although such systems may help ensure that inappropriate e-mails are deleted, provide good search facilities and fairly instant access to stored messages, and relieve the general burden on e-mail servers, there is little published information about their ability to ensure the management of e-mail messages for the long-term of fifty, a hundred, and two hundred years or more. Most are proprietary systems, which may make it difficult to access the messages if the supplier goes out of business and difficult to integrate the system with other digital object storage systems. Furthermore, use of such systems means that multiple locations must still be surveyed should disclosure be required, and multiple systems will have to be managed through time. Should institutions still decide to manage and preserve e-mail in separate systems, addressing these issues early on and making arrangements to counter them would be beneficial.

Avoiding fraudulent emails

Although e-mails can be used as evidence of transactions or communications, curators must be aware that it is possible to fraudulently create e-mail transmission files. These will generate false messages that can be used to 'prove' that data was sent when it in fact was not, that people were copied in on messages

²⁹ According to a *Special Briefing on E-mail Management* in AdLib magazine (March 2006), the average office worker handles some 75 emails per day.

when they were not, and that important data was submitted when actually it was obscured. The ease with which such messages can be created is a very strong argument for interception of messages at source or immediate archiving of messages by creators/recipients. Applying security protocols and persistent identifiers to message files at the point of ingest or receipt can help ensure that fraudulent e-mails do not permeate the collection.

3.3 Issues for Re-users of e-mail messages

Enabling the re-use of reliable data is one of the prime objectives in data curation. A data re-user is anyone who makes use of data, whether for academic research or learning, teaching or commercial purposes, legal investigation and auditing, or individuals petitioning access under the Freedom of Information and Data Protection Acts. Users of digital archives and collections will expect to be able to access, manipulate and analyse digital materials in ways that were never possible in the past with traditional paper-based materials.³⁰ However, re-use of reliable data is only possible if the data has been adequately curated.

Resource Discovery

Resource Discovery tools allow both curators and users to locate, retrieve, and use information in a large-scale environment. Tools must be provided by the curator to carry out these tasks and meet the needs of the data re-users, who often require software with 'intuitive user interfaces to facilitate and manage simple and complex information retrieval tasks'.³¹ E-mails should contain or link to resource discovery/description metadata that

enables such tools to identify and return materials – including not only e-mail content but also header information, attachments and other related objects - to data re-users. Such metadata will typically have been established when messages are first ingested into the repository and should cover attachments as well as the basic message. E-mail header information is a particularly useful source of resource discovery metadata.

Access rights management

As with any system, access restrictions and security must be implemented to ensure that the stored e-mails are protected from malicious alteration. These measures invariably extend to monitoring access to the messages and the audit trail mentioned earlier must feature again here. Individuals' privacy must be taken into account when providing access to stored messages, particularly in the context of the Data Protection Act. Curators may also wish to limit facilities for making copies of readily accessible messages, in which case some form of Digital Rights Management (DRM) technology may be appropriate.

Different access restrictions can apply to different parts of the digital collection; this should be addressed when developing access provisions. Finally, the legal restrictions described in section 2.4 must also be considered before stored messages are made available to other users.

Access provisions

Provisions must be made for providing access to stored messages and this may be accomplished through several different channels. For example, the Open Archival Information System (OAIS) Reference Model (described in section 4.5) does not grant access directly to the preserved material but by way of a Dissemination Information Package (DIP) that contains all of the information an end-user needs in an accessible format. The UK

³⁰ Seamus Ross, *Approaching Digital Preservation Holistically* – draft, not yet published (2006)

³¹ Renato Iannella, *The Resource Discovery Project*, as published in *Ariadne*
<http://www.ariadne.ac.uk/issue8/resource-discovery/>
(1997).

National Archives has a similar solution in place, whereby access is not provided directly to the preserved material but to a copy on the public access system, with an air-lock between that and the master system.³² Other approaches may generate access copies of the information on-demand,³³ or provide a copy of available components via an on-line web-server.³⁴

If access is available on-line, organisations may enable login-authorized access, for example via ATHENS or the Shibboleth Access management System or an on-line registration procedure.³⁵ The consequences of this should be considered before implementation: a lengthy registration procedure requiring personal details may dissuade the casual user from registering and accessing the resources on-line, however, it may be necessary for materials protected by copyright or containing restricted data.

Re-use requirements and activities

In many cases, the e-mails will be re-used for a different purpose than that for which they were originally created. In the short-term, e-mail content may be less likely to be re-purposed than attachment contents. Over the longer term,

the interactions displayed in a collection of e-mail messages may prove more interesting than the material contained in attachments that are, by that time, either out-dated or published in a final form elsewhere. The primary re-purposing of the materials is also likely to evolve as the material grows older: the historical and informational value of materials is likely to increase while the auditing or evidential value of the material decreases. The cultural and social value of the material may also increase over time (although this may occur in peaks and troughs). Large collections of e-mails will also have value for detecting data and patterns hidden within them.³⁶

Re-users may need to access data in different ways. Access to the preserved collection of a single individual may require access directly through an e-mail 'inbox' interface, in the same way that we access e-mail today. This allows a user to experience the e-mail collection in the same way as it was used, an important historical and social experience. Alternatively, users may wish to display the data in a different way, such as time-based or social network visualisation.³⁷ This enables re-users to analyse the collection through the rhythms it contains, particularly the changing relationships between the inbox 'owner' and his correspondents over time, and access visual representations of the owner's social e-mail network. Data re-users may require access to individual messages, or they may require messages from a range of individuals in an organisation's e-mail archive. Accessing the messages through an inbox-type interface is not necessarily the best – or most representative – way to achieve access in these circumstances, and alternative accessibility provisions will be required to meet the varying needs of the users.

³² For an overview of the access approach at the TNA, see *New Digital Archives at the National Archives*, http://www.nationalarchives.gov.uk/preservation/digitalarchive/pdf/project_background.pdf (undated).

³³ One of the earliest alternative preservation approaches to the basic migration option features migration on request – the generation of digital object in a new and contemporary format only when the object is requested – see *Migration - a CAMiLEON discussion paper* by Paul Wheatley, <http://www.ariadne.ac.uk/issue29/camileon/> (2001).

³⁴ This is the practice at Theater Instituut Nederland (TIN), see the ERPANET case study on TIN carried out in 2004 at http://www.erpanet.org/studies/docs/erpastudy_TIN.pdf.

³⁵ The Athens Access Management system provides users with single sign-on controlled access to web-based services. For more information on ATHENS, see <http://www.athens.ac.uk/>. Shibboleth is another single sign-on system, see <http://shibboleth.internet2.edu/> for more information.

³⁶ Judith Donath, *Visualising e-mail Archives – draft* http://smg.media.mit.edu/papers/Donath/EmailArchives_draft.pdf as quoted in Seamus Ross, *Approaching Digital Preservation Holistically*, op cit.

³⁷ Ibid

Users may also need access to contextual metadata that explains what a resource is about so that they can re-use the data in a meaningful

way and place it in context with other parts of the collection. This can be provided as part of the overall record 'package'.

4. Preserving E-mails

4.1 Options for preserving e-mail messages

Print to paper

Simple printing to paper was the singularly most prolific approach for e-mail preservation until the late 1990's. Opinions on the suitability of this approach changed with the U.S. cases of *Armstrong v. Executive Office of the President* and *Public Citizen v. Carlin*. The *Armstrong* case, which lasted from 1989 to 1996, resulted in then-U.S. Archivist John Carlin issuing revised regulations and guidelines in 1995 for the management of electronic records and e-mails, including General Records Schedule 20 (GRS 20). Public Citizen (a self-declared non-profit public interest organisation) and other historical and library organisations took issue with these guidelines and launched the case of *Public Citizen v. Carlin* in 1996. Public Citizen challenged the Archivist's issuance of GRS 20 because it allowed agencies to delete electronic mail and electronic word processing files from 'live' systems once a copy had been made and preserved in either a paper or electronic record-keeping system, which Public Citizen considered unacceptable.

GRS 20 allowed the replacement of electronic mail files with printouts of the e-mails as long as the printouts contained relevant header information (including, for example, the names of all recipients of an e-mail). This was the Archivist's interpretation of the *Armstrong* verdict, in which relevant header information was deemed to be part of the e-mail record and thus must be preserved. The District Court ruled in 1997 that the Archivist had exceeded his authority in issuing a general records schedule covering all e-mail records, and stated that deletion was therefore unauthorised in law until such time as each agency separately scheduled its programmatic e-mail records. An appeal was immediately launched and in 1999

the U.S. Court of Appeals for the District of Columbia granted the appeal and reversed the 1997 decision. The result of this is that electronic versions of e-mail records may be deleted provided that those records are stored in either paper or electronic files, and provided that the e-mails in hard copy contain all relevant transmission and receipt data.³⁸

Despite the approved legality of the approach, this case was fundamental in changing attitudes towards e-mail preservation. In 1999, the US National Archives and Records Administration (NARA) launched its Collection Based Long Term Preservation research initiative to develop a solution to preserve records, including e-mails, in electronic form.³⁹ Many other bodies now also disapprove of the print-to-paper approach, the reasons for which are numerous:

- The printed version of the e-mail will not always contain all of the same information as the electronic one;
- Printed records lose their functionality;
- Printed records cannot be consulted simultaneously;
- Active links to attachments and embedded files are broken;
- Some organisations require that records be kept in their original formats to verify authenticity;
- Storing printed records takes up expensive physical space; as e-mail is

³⁸ *E-mail, Laws, and Backup Tapes: How can my agency cope?* Jason Baron, 2004.

<http://www.armamar.org/nova/Downloads/NARA%20E-mail%20Forum%202004.ppt>. See also Jason R. Baron, "The PROFS Decade: NARA, E-mail and the Courts," chap. 6 in Bruce Ambacher, ed., *Thirty Years of Electronic Records* (Scarecrow Press 2003).

³⁹ More information about the SDSC/NARA Collection-Based Long Term Preservation initiative is available from <http://www.sdsc.edu/NARA/>.

supplemental to paper and does not replace it, the number of e-mails that must be preserved exceeds traditional storage requirements;

- Further costs are associated with printing such as printers, ink, and paper;
- Rich searching capabilities are lost
- E-mail is electronic by its very name and nature;
- Digital signatures can be meaningless when printed out to paper (although preservation of digital signatures is another issue).

Transferring records from one medium to another has a long-standing precedent in the replacement of paper-based records with microfilm versions. The difference however, is that whilst paper-based records are inherently spatially fixed and thus can be fully replicated on microfilm, digital records may have inherent functionality that cannot be adequately represented with 2D materials. The potential characteristics or attributes of the records preclude their transfer onto an alternative carrier. Nonetheless, many organisations continue to implement this approach: as recently as 2003, the Council of Europe reported in an ERPANET case study that:

'[f]or certain categories of records, print-to-paper is the only means of preservation available. This holds especially true for e-mail, where no possibility of digital archiving exists...' [within the organisation]'.⁴⁰

Retain within e-mail systems

The time-honoured 'do-nothing' approach of leaving e-mails in user's e-mail accounts and relying on system back-ups or hoping that individual users will manage their own mail is

not a viable solution. Messages are often in a proprietary format, are inaccessible to the wider work group, and devoid of links to the essential functionality that a document management system (DMS) or RMS can provide and link messages to a business process and a retention schedule. This approach is, by and large, the reason why e-mail mis-management has become such an issue. Inappropriate messages languish in inboxes or on back-up tapes, only to become evidence in legal cases brought by disgruntled employees or regulatory bodies and result in financial penalties for the organisation concerned. Similarly, e-mails that ought to have been retained but which are not found in an initial sweep of the e-mail system can result in expensive recovery procedures.⁴¹ E-mail cultural heritage, i.e. correspondence between notable literary figures, may be conspicuous by its absence in the twenty-second century if contemporary personal e-mail systems belonging to such figures are not efficiently managed. In these cases, where messages are downloaded using POP3 and stored on a user's home PC, personal but nationally and culturally valuable correspondence can easily be lost with the 'do-nothing' approach, through something as simple as a infected or burnt-out PC with no e-mail back-up.

Convert to standards

By far the most popular current approach for long-term management and curation of e-mail messages involves migration or conversion to standards. The chosen standard must be suitable and appropriate for e-mail preservation, and able to meet authenticity and preservation requirements. As different types of objects have different preservation requirements, so different standards are

⁴⁰ ERPANET case study on the Council of Europe, see http://www.erpanet.org/studies/docs/erpaStudy_COE.pdf. Other institutions reported similarly.

⁴¹ In addition to those cited in the introduction, several more cases are cited in an article by Melissa Campbell in the Alaska Journal of Commerce, April 2005. See http://www.alaskajournal.com/stories/120405/hom_2005_1204011.shtml.

suitable for different types of digital objects. The standards most relevant for e-mail preservation are PDF, TIFF, RFC 2822, and XML.

PDF and TIFF can be used to store an image of the e-mail message, thereby capturing the message content, appearance and on-screen structure as perceived by the creator/recipient. The manner in which an e-mail can be saved to PDF and TIFF will depend on the facilities available within a given computing environment. Whilst the content of the message will then be readily accessible using an image viewer, the complete header metadata may not be preserved without additional efforts and files. Procedures must be devised to maintain the links between the extra files. Attachments must be extracted from the message to undergo additional preservation action; again, these must be linked to the message file and the metadata. Ultimately, although PDF and TIFF are suitable for some types of digital record preservation, they are not ideally suited towards the preservation of e-mails and certain other record-types. For example, in 2004 the US Court declared TIFF files to be insufficient for documents in a class action case as they did not contain all of the metadata from the original file and it was not how they were stored in the Defendants' usual course of business.⁴² This argument against using the TIFF standard on documents could easily be extended to apply to e-mails.

RFC 2822 is the standard format for the original e-mail transmission file. As a source file, it contains all of the data that must be preserved in order for the e-mail message to be authentic. It is ideal as a starting point for a preservation format, but alone it is not enough. The main problem with simply preserving this file is the encoding of attachments.

⁴² *In re Verisign, Inc. Sec. Litig.*, 2004 WL 2445243 (N.D. Cal. Mar. 10, 2004), as cited in Ross, *Approaching Digital Preservation Holistically*, op cit.

Attachments might include word processing documents, spreadsheets, other e-mails, even pieces of software. These are all binary files. Since e-mail transmission systems were designed to handle only plain text, binary files must be converted to ASCII before they can be sent through the e-mail system and this function is performed by the sending and receiving e-mail clients. Simply storing the RFC 2822 file direct from the client therefore means that attachments will be encoded. This is not ideal for preservation and the files must either be decoded before entering long term storage, or a decoding mechanism also stored and preserved. If attachments are extracted and saved separately, persistent linkages between the related files must be implemented, as with PDF or TIFF files (above). Finally, a mechanism for rendering the RFC 2822 files into messages with appropriate formatting must be created. If these issues are addressed, storage of the RFC 2822 files is the simplest manner to achieve enduring e-mails. Note however that although the *.eml extension is the most widely used extension for files in this format, not all e-mail applications offer this nor allow easy access to the RFC 2822 files.

Using XML to mark-up the messages is a popular alternative. XML is often hailed as a magic bullet and it has many useful implementations for preserving e-mails.⁴³ Conversion software can be developed and deployed to mark-up the message content and the header metadata in a single XML file. The XML file is processed and rendered using a related member of the XML family, XSLT, to render the formatted message on-screen. Separate preservation action is required for attachments, appropriate to the attachment

⁴³ See for example Maureen Potter's presentation on XML for Preservation at the ERPANET workshop on XML as a preservation Strategy in Urbino, 2002, available at http://www.erpanet.org/events/2002/urbino/presentations/Testbed_Erpanet_XML.pdf.

record type (for example, spreadsheets). In some cases this may also be XML. XML can further be used to retain the links between the XML message file and attachments. The XML can be parsed to make sure that it is suitable for long-term preservation and that it meets defined schemas. The advantage of using a single family of standards to preserve the e-mail is obvious: if migration becomes necessary at a later date, the migration procedure will be that much simpler as the source files are consistently saved using the same mark-up language. Future migration can be even more straightforward if the files are marked-up in a manner that makes the content human-understandable, meaning they can remain accessible in that format over periods of hundreds of years. Furthermore, new content files can be produced in several different formats from the original XML files, which is advantageous for users who may require broad format access to the message content.⁴⁴ The Belgian DAVID project in its report on Archiving e-mails rightly makes the point that XML files will only remain useful as long as computers can read Unicode characters; however, it is generally assumed that Unicode will remain the basis for new character tables.⁴⁵

The Digital Curation Centre website offers access to a range of non-technical information about available and emerging standards and specifications that facilitate the electronic exchange and management of digital information. This information was originally

collected and developed as part of the DIFFUSE (Dissemination of InFormal and Formal Useful Specifications and Experiences) project. The DIFFUSE project ended in 2003 and the collection is now hosted and maintained by the DCC.⁴⁶

4.2 Dealing with Attachments

The importance of maintaining links between attachments and message content and of identifying a procedure to decode attachments out of their transmission coding is stressed in the sections above. The issue of **preserving** those attachments requires further attention and is not always straightforward. Attachments may comprise text documents, graphics, spreadsheets, video and audio files, presentations, web-pages, compressed or encoded files (such as *.zip or *.tar files) which contain further files, executables, and other types of data files such as statistical analysis files. This is to name but a few. The format of the files is potentially unlimited: hundreds of file formats are listed on various Internet sites dealing with file format information, and additional file formats are constantly being developed.⁴⁷

This range of formats makes it impossible to develop a 'one-size-fits-all' approach for preserving attachments. Developing a more extensive preservation strategy to cater for different file and record types across a whole organisation can establish solutions for the most commonly used formats.⁴⁸ An e-mail

⁴⁴ See *Many Outputs — Many Inputs: XML for Publishers and E-book Designers*, Hillesund in the Journal of Digital Information, Volume 3 Issue 1 Article No. 101, 2002-08-06

<http://jodi.tamu.edu/Articles/v03/i01/Hillesund/> and *XML: One Input — Many Outputs: a response to Hillesund*, Walsh, in Journal of Digital Information, Volume 3 Issue 1 Article No. 165, 2002-09-12 <http://jodi.tamu.edu/Articles/v03/i01/Walsh/>.

⁴⁵ *Archiving e-mails*, Boudrez F & Eynde, Sofie Van den, op cit.

⁴⁶ See the DIFFUSE collection on the DCC website at <http://www.dcc.ac.uk/diffuse/>.

⁴⁷ Websites such as <http://www.wotsit.org/> and <http://whatis.techtarget.com/fileFormatA/0,289933,sid9,00.html> aim to provide information or specifications of file formats. Over 3000 formats are currently identified in the WhatIs website.

⁴⁸ For example, the Belgian eDavid Expertise Centrum has issued advice on its preferred formats for storing attachments. Recommended formats for text documents

preservation strategy may then be integrated with this overall approach and the relevant preservation action identified and taken for attachments in different formats.

In many cases, attachments must be preserved alongside the e-mail message and the link between the files maintained to ensure the provenance and integrity of the record. This link between preserving e-mail messages and preserving attachments comprising other record types is a further reason for integrating e-mail archiving and management into a wider organisational strategy, rather than establishing a separate e-mail storage facility.

4.3 Dealing with Digital Signatures

A digital signature establishes the identity of the person who sent an e-mail message. It also confirms that the content of the e-mail has not been altered since it was signed by the sender and prevents the sender of the message from denying that they sent it (non-repudiation). Put simply, a signature file is appended to an e-mail by the sender, and the recipient verifies the sender's identity by accessing the signature details (or certificate) and checking the sender's identity via the company or person who issued the certificate.⁴⁹ The EC Directive on Electronic Signatures gives digital signatures in the Member States the same legal status as

are TIFF and the Open Document Text Document format ODT; for spreadsheets the Expertise Centrum recommends TIFF, XML and ODS (the Open Document Spreadsheet format), and for audio files, the Waveform audio format WAV. Further recommendations can be found in the report 'Filing and Archiving Email' (2006), op cit.

⁴⁹Although the terms 'digital signature' and 'electronic signature' are frequently used interchangeably, there is a difference between the two. A digital signature, described above, uses PKI technology to guarantee data integrity and non-repudiation of transactions. An electronic signature is an electronic image that is physically or logically attached to the signed data and is commonly biometric.

handwritten signatures, as long as certain technical specifications are met.⁵⁰

For the signature to retain value over time, the signature must be verified as valid at the time of receipt as it is unlikely that the act of verification will continue to be possible over the long-term. Certification providers or authorities are not required to preserve data for extended periods after certificates have expired. Furthermore, the contents of the message are effectively 'altered' by migration to a new preservation format and this invalidates the signature.

4.4 Preservation Metadata

Metadata allows a digital object to be meaningfully managed, preserved, and used. A number of international projects or initiatives have focused exclusively on preservation metadata, i.e. metadata specifically required to support long-term preservation of digital objects. An excellent overview of the issues surrounding a range of metadata types can be found in Michael Day's Metadata instalment of the Digital Curation manual, so will not be covered in great detail here.⁵¹ Suffice to say that metadata has multiple uses and functions in a preservation environment: metadata can be used to record the validity of a signature at the time of receipt, can be used to maintain an audit trail, to determine and verify authenticity and integrity, and is a major component in the OAIS Information Packages to document the object at the point of ingest, provide

⁵⁰Directive 1999/93/EC of the European Parliament and of the Council of 13 December 1999 on a Community framework for electronic signatures
http://europa.eu.int/eur-lex/pri/en/oj/dat/2000/l_013/l_01320000119en00120020.pdf.

⁵¹Day, Michael (November 2005) *Metadata* DCC Digital Curation Manual Instalment
<http://www.dcc.ac.uk/resource/curation-manual/chapters/metadata/>.

preservation and access services, and allow the re-user to understand the object at a later date. Metadata thus plays a key role in the preservation of authentic digital records.

4.5 Long term storage and archiving

Regardless of the selected preservation strategy, a storage infrastructure is also required to manage and maintain the e-mail messages through time. For this purpose, institutions should consider implementing a digital repository or digital archive. Digital repositories offer a convenient infrastructure through which to store, manage, re-use and curate digital materials. The meaning of the term 'digital repository' is widely debated. Contemporary understanding has broadened from an initial focus on software systems to a wider and overall commitment to the stewardship of digital materials; this requires not just software and hardware, but also policies, processes, services, and people, as well as content and metadata. Repositories must be sustainable, trusted, well-supported and well-managed in order to function properly. A digital archive has the same essential characteristics but with an additional and express commitment to preserve the integrity of the contents for the long-term, and provide for long-term storage and access to its contents.⁵²

The selected storage facility should be able to manage both content and metadata, and preferably content objects comprised of more than one file. 'Encapsulation' is a way to store message files, metadata, and the source transmission file together in the same bundle. It

⁵²The seminal 1996 report from the RLG Task Force on Archiving Digital Information, *Preserving Digital Information: Final Report & Recommendations* contains an interesting exploration of the term 'digital archive' and the difference between a digital archive and a digital library. See <http://www.rlg.org/ArchTF/>.

is most effective when used in conjunction with a conversion to standards approach. It has been proposed by several individuals or projects, notably Jeff Rothenberg,⁵³ Michael Day,⁵⁴ the Digital Preservation Testbed,⁵⁵ and Thom Shepard.⁵⁶

Encapsulation is also a fundamental aspect of the widely acclaimed Open Archival Information System (OAIS), a reference model for archival management and storage. The OAIS model was originally developed by the Consultative Committee for Space Data Systems (CCSDS) for use with space data and has been adopted by numerous private and public bodies for a range of data types since approval as an ISO standard in 2003.⁵⁷ ⁵⁸ Data within the OAIS reference model is contained in a series of Information Packages, namely a Submission Information Package (SIP), an Archival Information Package (AIP) and a Dissemination Information Package (DIP). Most relevant for the purposes of preservation is the Archival Information Package (AIP). This contains the information that is the focus of preservation, along with any metadata required to support the OAIS services. The

⁵³Jeff Rothenberg, *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation* (1999)

<http://www.clir.org/pubs/abstract/pub77.html>.

⁵⁴Michael Day, *Metadata for Preservation*, (1998).

<http://www.ukoln.ac.uk/metadata/cedars/AIW01.html>.

⁵⁵Digital Preservation Testbed, op cit.

⁵⁶*The Universal Preservation Format (UPF):*

Conceptual Framework, Thom Shepard, 1998

<http://www.rlg.org/preserv/diginews/diginews2-6.html#upf>

⁵⁷*Open Archival Information System (OAIS) Reference Model* Developed by the CCSDS and published as an ISO Standard in 2003.

<http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=24683&ICS1=49&ICS2=140&ICS3>

⁵⁸A concise introduction to OAIS is presented in the DPC Technology Watch report written by Brian F Lavoie, *The Open Archival Information System Reference Model: Introductory Guide* (2004) http://www.dpconline.org/docs/lavoie_OAIS.pdf.

reference model is conceptual and does not offer a technical solution which can be directly implemented; however, the functions and processes of an OAIS can be found in and mapped against many current digital repository or archiving models.

Once in long-term storage, a technology watch must be implemented to ensure that stored objects remain accessible and authentic despite changes in the technological environment, and security protocols must be implemented to protect the stored records against unwarranted intrusions and unauthorised access. The ISO 9000 family of standards for quality management can facilitate this, as can ISO 27001 for information security management, and the institutional records management standard ISO 15489.⁵⁹ The DCC is currently working with the US Research Libraries Group (RLG) towards implementation of an Audit and Certification framework that addresses these and other trust issues in digital archives.⁶⁰

⁵⁹Further information and purchasing details for all of these standards can be found on the ISO website, <http://www.iso.org/>

⁶⁰For more information on the Audit and Certification Programme, see the RLG paper *'Trusted Digital Repositories: Attributes and Responsibilities'* (May 2002) <http://www.rlg.org/legacy/longterm/repositories.pdf>. and Ross, S & McHugh, A *'Audit and Certification of Digital Repositories: Creating a Mandate for the Digital Curation Centre (DCC)'* (October 2005) http://www.rlg.org/en/page.php?Page_ID=20793#article
1.

5. Practical Steps

Organisations with a responsibility to curate e-mails should develop an approach that caters for the different stages of the e-mail life cycle, from creation, active use, archiving, preservation, access and re-use, to disposal or transfer of stewardship. The risks of not developing an approach to curating e-mail and e-mail collections should be measured in an appropriate risk assessment; this will help secure funding and management backing to achieve the task at hand.

The issues addressed throughout this report should all be considered when developing an approach towards curating e-mails. The first step is to gain management buy-in, which may be facilitated by the results of an information compliance audit and risk assessment. Backing and involvement of other stakeholders is also important, particularly in terms of co-operation and communication between records managers, archivists, IT staff, and departmental 'heroes' who can enhance the visibility of the curation strategy amongst users and data creators. Subsequent activities must address, at the very least, the following three areas:

- Policy
- Education & Training
- Capture & Preservation

5.1 Developing an e-mail policy

Policy indicates the extent to which activities are embedded in an organisation and the importance attributed to them. Organisations with a responsibility to curate e-mail messages and other records must establish the importance and relevance of such activities in a policy document. The particular type of policy document may vary, according to the needs of the institution. It is possible that more than one policy will be required and created, for

example, the computing department may deliver a policy to address acceptable use, and the records management/archives department produce a policy addressing record-keeping activities for e-mails, or integrate e-mails into their existing policy framework. Policies must be integrated into the institution's overall policy framework to ensure wide-reaching impact and credibility.

In keeping with the life-cycle model, the following topics should be addressed. Elaboration at this point is not generally required and further details can be provided elsewhere, for example, in retention schedules or alternative policy documents:

- The need for appropriate creation practices;
- Using business e-mail infrastructure for personal mails
- Using personal e-mail accounts for business mails
- Shared responsibilities for managing e-mails
- Shared access to e-mail messages
- Legal obligations and opportunities for training to meet them.
- E-mail retention within overall records management activities
- E-mail preservation within overall preservation management activities
- Disposal of inappropriate or unnecessary messages

Integrating e-mail management into a wider records management approach ensures that electronic records of a particular topic can be kept together and remain widely accessible, whilst clear policies help ensure conformance with institutional requirements.

5.2 Educating users and stakeholders on their responsibilities in e-mail curation

Complete conformance can only be achieved if training is provided. Appropriate staff should develop and provide training and advice for all users with e-mail management and preservation responsibilities. The following list provides a useful starting point of issues on which user guidance should be provided:

When composing an e-mail -

- Which format to send mail in (plain text/html/rtf)
 - Using a template to ensure that e-mails comply with organisational house-styles
 - Entering a meaningful subject line
 - Using the address book to capture full names of recipients
 - Using distribution lists and how to capture the names of members
 - To BCC or not to BCC – BCC details cannot be harvested from received messages as they are stripped out by the servers
 - When to incorporate attachments, embedded items, and links, and when to put the content directly in the e-mail message body
 - Using formal language – formal e-mails should be approached as formal letters: don't put anything in an e-mail that you wouldn't put in a letter.
 - Using message threads (i.e. running conversations between two or more parties by way of the 'reply' or 'reply all' button) in a manner that will make it possible to decipher the contents in the future. New content should be clearly identified from the original content:
1. Consistency - be consistent in where the replied text is inserted into a message, either above or below the original. Do not mix the two, as this makes it

difficult to follow the thread of the message at a later date;

2. Enter replies in a solid block, rather than inserting new content at various points in amongst the original content. This is summarised as 'block, don't quote' – for the same reasons as above;
3. Exercise caution when using message flags as although they will be included in the transmission file, your recipient's e-mail application may not display them.
 - Retaining copies of appropriate 'sent' mails and deleting non-required copies of sent mails.
 - Using a signature block so that recipients can contact you by means other than e-mail, and to clearly identify your position and organisation.

When addressing user storage of messages -

- Which storage format to use
- Deleting transient emails
- Using a folder system that is consistent with the organisation's record-keeping system
- Regular submission of relevant emails to the selected record-keeping system, if no automated procedure exists. This may be a organisation-wide electronic record-keeping system with built in retention scheduling and metadata fields, or a shared server in a folder system.
- Ensuring attachments are stored with e-mails and that the link between the two is not broken.
- Applying these principles to messages in the 'sent' items folder, as well as received messages.

In addition to the above, training programmes should cover legal responsibilities, particularly with reference to Data Protection Act and the Freedom of Information Act. Staff must be

informed of potential monitoring activities, in light of the role of e-mail messages in business processes. Compliance advice for the RIP Act is available from the JISC.⁶¹

Records Management/Curators and IT staff must communicate and be aware of each others responsibilities so that viable and compatible solutions can be implemented: there is no point in developing and implementing an e-mail retention schedule if IT policy is to delete inbox contents after a sixty-day period to ensure staff do not exceed their storage quotas. An overview of some current policies that address responsibilities and creation practices can be found in Appendix 1.

5.3 Implementing a solution to capture and preserve e-mails for specific retention periods

Messages that have long retention periods (i.e. anything over ten years) should commonly be migrated to a standard format to ensure their persistence and authenticity over time. Migration to a widely accessible format should also be considered for messages with shorter retention periods (e.g. five years) that use closed, proprietary formats, so as to ensure their accessibility over this period independently from a specific e-mail client. This is not a cut-and-dry rule and organisations must assess their own organisation and the resources available before ultimately deciding upon a particular strategy.

Capture and preservation strategies are best developed by collaborative groups comprising at the very least representatives from Creators, Management, Records Management/Curators and IT. Capture should ideally take place as close to the point of creation/receipt as possible. This increases the quality of metadata

applied to the message and enables colleagues and other appropriate parties to access the message whilst still in active use. Appropriate preservation solutions will depend upon organisational requirements and resources: whilst an on-site, integrated solution is best for some, off-site and externally managed solutions may be best for others.

Technically, a preservation approach that utilises the multi-faceted benefits of XML is compatible with current and cutting-edge approaches to preserving e-mails. However, such an approach takes time to develop and e-mail preservation should not be delayed until the optimum technical infrastructure has been developed. Any such delay increases the risk of damage to the integrity and authenticity of the e-mails, and is likely to result in the loss of some messages. An interim solution should be implemented, for example, storage of RFC 2822 files with links to attachments in suitable formats, integrated into the organisational digital records collection system, using storage media with a long shelf-life that is refreshed at regular intervals until the chosen system is finalised.

Security protocols should be implemented to ensure that the preserved messages are safe from interference and unauthorised access. Approaches should also be reviewed at regular intervals to ensure that they remain satisfactory to the organisation's requirements and that they are successful in preserving the chosen materials.

⁶¹For JISC Compliance advice for the RIP Act, see http://www.jisc.ac.uk/index.cfm?name=pub_smbp_ripa.

6. Future Developments

The proper and active curation of e-mail messages, as with websites, is still in a period of immaturity. Most organisations are still focused on trying to bring their e-mail management to order and have not yet begun to practically address archiving and preservation issues. This is despite the emergence and enthusiastic reception of XML-based archiving and preservation solutions from archiving institutions. As the number of legal cases involving failure to adequately manage and preserve emails increases and spreads from the US to the UK, we can expect more and more commercial e-mail archiving solutions to appear on the market. Whether such commercial solutions can truly offer a complete and long-term solution has yet to be determined.

At the user-end of the curation process, client-based e-mail management is set to change. Google Mail (commonly and hereafter referred to as Gmail) is leading the way, offering users a web-based e-mail system with currently over 2.7 GB of free storage.⁶² However, the phenomenal amount of free storage space is not the real revolutionary aspect; that lies instead in the way Gmail approaches e-mail organisation. E-mails are grouped and presented according to their threads, rather than the classic chronological presentation method, and Gmail allows users to classify threads by allocating 'labels' that appear directly beside the thread title.

Google considers Gmail's enhanced search capabilities to be at the heart of its success. The search engine does, of course, utilise Google technology. The service is advertised with the benefits that:

'you no longer need to set up folders, file your mail, or remember where you stored your messages. Just search for what you want. You'll not only find the message you have in mind, but all the other messages that are part of the same conversation – arranged in chronological order so you can easily put everything in context.'

This alternative form of message display makes it easy to follow the thread of a conversation without affecting the ordering of an entire folder or inbox. Furthermore, users have the option of developing and applying 'labels' to conversations, as another way of organising them and to facilitate searching. Traditional 'filing' of messages into folders and sub-folders is no longer necessary. Of course, it has yet to be seen how this approach will translate into the domain of professional e-mail systems as it is currently mostly utilised for personal use. Despite this, the popularity of Gmail indicates that Google certainly aren't onto a loser, and wide levels of private user-take up are often indicative of commercial success (albeit somewhat adapted) at a later stage.

Enhanced search capabilities are also at the heart of the forthcoming Windows VISTA operating system, whose release package features the new 'Windows Mail' application.⁶³ The application utilises a built-in 'Quick Search' function that can search not only the contents of the e-mail folders but throughout the entire operating system when you search from Search Explorer. This enables users to search, and access, e-mail messages and other types of documents within a single search mechanism and supports a more integrated

⁶²Gmail: <http://mail.google.com/>. The amount of storage space available to users has continually increased since the service was launched.

⁶³Windows VISTA:
<http://www.microsoft.com/windowsvista/>.

approach to e-mail use and management.

Another approach that may prove popular in the near future is the use of e-mail visualisation applications. Recent (although by no means the first) research in this area by Fernanda Viégas et al has resulted in the development of two visualisation tools, Post History and Social Network Fragments, which provide users with an alternative approach to accessing and understanding e-mail collections by presenting 'visualisations' of e-mail relationships and social landscapes.⁶⁴ The user can then explore an entire e-mail 'world' rather than simply accessing and reading separate messages. The technology enables users to uncover social patterns within e-mail archives and this significantly enhances the potential re-use value of a collection. The current emphasis and derived value is oriented on the e-mail archives of a single person and scaling up to represent an archival collection would require extensive re-orientation of the approach to ensure that the visualisations were still meaningful. Developers have also uncovered a more immediate value of the application for real-time use and management of personal e-mail collections, although it is not immediately clear how this would extend to an office environment. Similar work in this area by Perer, Schneiderman and Oard is also notable.⁶⁵

Considering the future of e-mails in the wider

⁶⁴Fernanda B. Viégas et al, *Digital Artifacts for Remembering and Storytelling: PostHistory and Social Network Fragments*, as published in the Proceedings of the 37th Hawaii International Conference on System Sciences (2004) http://alumni.media.mit.edu/~fviegas/papers/posthistory_snf.pdf.

⁶⁵Perer, Schneiderman and Oard 'Using Rhythms of Relationships to Understand E-mail Archives' (2005). The team developed a useful classification that delineates types of interactions with e-mail collections in their paper. The classification system can also be applied to creation and preservation stages of e-mail curation. See <http://hcil.cs.umd.edu/trs/2005-08/2005-08.pdf>

context of the immediate computing environment, the ongoing evolution of software capabilities may actually lead to a reduction in the extent to which e-mail is employed as a records communications tool. The prolific use of e-mail for a whole range of things that were probably never envisaged when e-mail was originally conceived, from an instant messaging tool to file transfer, storage, and even project management, along with the growing number of listserv messages, SPAM, and flagrant (mis)use of the CC and BCC functions, all contribute to the growing number of e-mails sent and received on a daily basis and compound issues of curation, management, and preservation.⁶⁶ The appearance of new technologies that offer a more appropriate channel for some of the ways in which e-mail is used may or may not address the problem.

Instant Messaging (IM) is the forerunner in this challenge to e-mail. IM is certainly the preferred method of communication between teenagers, who perceive e-mail as a tool for communicating with institutions and other formal relationships.⁶⁷ Usage in the workplace is growing, especially between colleagues who work remotely, and IM is a particularly useful alternative to e-mail when a protracted exchange is required to address a problem in a short space of time.⁶⁸ Its increasing use as a workplace tool is visible in a growing number of documents addressing the need to manage IM as potential records.⁶⁹ RSS is the second place challenger, allowing an alternative option

⁶⁶The Register, *What's the Future of E-mail?* (September 2005) http://www.theregister.co.uk/2005/09/21/e-mail_future/.

⁶⁷The Register, *E-mail? That's for old farts!* (July 2005) http://www.theregister.co.uk/2005/07/28/e-mail_for_old_farts/.

⁶⁸Greenspan, Robyn, *Workplace IM Showing Growth* (September 2004) in ClickZ Stats Professional <http://www.clickz.com/stats/sectors/professional/article.php/3402961>.

⁶⁹*Why Instant Messaging Management?* (May 2004) the ePolicy Institute <http://www.epolicyinstitute.com/imr/intro.pdf>.

to e-mails for making announcements. RSS (Really Simple Syndication/RDF Site Summary) is a news feed technology that allows users flexibility in managing the information they receive. One of the most useful aspects of RSS insofar as e-mail curation is concerned, is the separation of e-mail messages that may be records from news announcements that are not. The inbox is no longer filled with news announcements, and is therefore that little bit less cluttered and easier to manage. Although this could also largely be achieved by developing and implementing filters, many users appear reluctant to do so for fear of losing 'valuable' messages in the filtering system. RSS feeds can be received via some e-mail clients (for example, Mozilla Thunderbird), or through browsers.

Wikis, discussion forums, blogs, and Voice over Internet Protocol (VoIP) also offer electronic alternatives to using e-mail. No doubt the number of available technologies will increase with time. However, it is unlikely that e-mail will die out completely – it has simply too many uses within a single environment to become redundant. It is more likely that the e-mail environment, the way in which users manage communications, and how they integrate them into the management of other records and document types, will evolve. E-mail won't 'die', not for a long time yet, but it almost certainly will change.⁷⁰

What will this mean for future activities in the preservation and archiving of e-mail messages? At the time of writing there is very little further research taking place specifically into long-term curation of e-mail messages. The few programmes that are concerned with this, or which have been in the past decade, are discussed in Appendix 2. This is to an extent reflective of the fact that there are, as yet, also

very few institutions who have moved beyond the e-mail management stage into practical archiving and preservation of e-mail messages.

⁷⁰Brian Kelly *Web Focus: Must E-mail Die?* In Ariadne, Issue 45 (October 2005)
<http://www.ariadne.ac.uk/issue45/web-focus/>.

7. Conclusion

E-mail curation is a many-layered thing. To date, most institutional activities have focussed 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 organisational and cultural challenges remain a significant barrier. This is the case not just for e-mails, but also for institutional records management generally, which has only in the past few years truly begun to embrace possibilities for electronic archiving of electronic records.

Relying on the 'print-to-paper' option as a primary preservation strategy is unsatisfactory for long-term preservation of an electronic communications medium. It results largely from a lack of digital experience in institutions with a responsibility to manage and preserve

digital records including e-mails, a lack of trust in intangible digital objects, and a lack of recognition regarding the levels of co-operation between different stakeholder parties that is necessary to ensure the persistence of digital records. Digital records management is complex, but that is no reason to rely instead on print-outs. As technical solutions develop, trust will grow and experience will be gained. The final key to the solution is the development of collaborative relationships between all parties with responsibilities for creation, management, and preservation of e-mails and digital records generally. Given the complexity of the relationships and the interdependency between the activities for which different groups have responsibility, it is clear that only through collaboration can a successful solution be implemented to manage e-mails effectively across their entire life-cycle, however long (or short) that may be.

Glossary

A full glossary of terms is available from the DCC website:

<http://www.dcc.ac.uk/resource/glossary/>

References

- Allman, Thomas Y (Sept/Oct 2005) *E-mail Retention: Time for a New Approach* AIIM E-Doc Magazine
http://www.edocmagazine.com/quick_article.asp?ID=30580 [Accessed 02 June 2006]
- Baron, Jason (2003) *The PROFS Decade: NARA, E-mail and the Courts*, chap. 6 in Bruce Ambacher, ed., *Thirty Years of Electronic Records* (Scarecrow Press 2003).
- Baron, Jason (2004) *All the President's E-mail: Electronic Recordkeeping Policies and Practices in the Executive Office of the President*, ERPANET Workshop on Audit & Certification in Digital Preservation, 2004
http://www.erpanet.org/events/2004/antwerpen/presentations/erpaWorkshop-Antwerpen_Baron.ppt [Accessed 02 June 2006]
- Baron, Jason (March 2004) *E-mail, Laws, and Backup Tapes: How can my agency cope? Cautionary Tales from the Archives*, NARA E-Records Forum, 2004
<http://www.armamar.org/nova/Downloads/NARA%20Email%20Forum%202004.ppt> [Accessed 02 June 2006]
- Blanchette, Jean Francois (2004) *The Digital Signature Dilemma in Annals of Telecommunications* (preprint)
<http://polaris.gseis.ucla.edu/blanchette/papers/annals.pdf> [Accessed 02 June 2006]
- Bloor, Robin (July 2005) *What's the future of email?* The Register
http://www.theregister.co.uk/2005/09/21/email_future/ [Accessed 02 June 2006]
- Boudrez F & Eynde, Sofie Van den (August 2002) *Archiving E-mail*
<http://www.expertisecentrumdavid.be/davidproject/teksten/Rapporten/Report4.pdf> [Accessed 02 June 2006]
- Boudrez, Filip (2005) *Digital containers for shipment into the future*
http://www.expertisecentrumdavid.be/docs/digital_containers.pdf [Accessed 02 June 2006]
- Boudrez, Filip (2006) *Filing and Archiving E-mail*
http://www.expertisecentrumdavid.be/docs/filingArchiving_email.pdf [Accessed 02 June 2006]
- Campbell, Melissa (December 2005) *E-Mail may come back to haunt you* Alaska Journal of Commerce
http://www.alaskajournal.com/stories/120405/home_20051204011.shtml [Accessed 02 June 2006]
- Coehn Daniel J (Summer 2005) *The Future of Preserving the Past CRM: The Journal of Heritage Stewardship* Vol 2 02 pp 6-19
http://www.cr.nps.gov/crdi/publications/CRM_Vol_2_02_Viewpoint.pdf [Accessed 02 June 2006]
- Cole, Richard & Eklund, Peter (September 1999) *Analyzing an E-mail Collection Using Formal Concept Analysis* in Principles of Data Mining and Knowledge Discovery: Third European Conference PKDD'99, Prague (Proceedings), published by Springer-Verlag GmbH
- Cooper, John & Bowles, Philip (August 2003) *Evidential Value of E-mail* Datassec White Paper
<http://www.datassec.co.uk/whitepapers/The%20Evidential%20Value%20of%20email.pdf> [Accessed 02 June 2006]
- Day, Michael (August 1998) *Metadata for Preservation* CEDARS Project Document AIW01
<http://www.ukoln.ac.uk/metadata/cedars/AIW01.html> [Accessed 02 June 2006]
- Day, Michael (November 2005) *DCC Manual Instalment: Metadata* (DCC)
<http://www.dcc.ac.uk/resource/curation-manual/chapters/metadata/> [Accessed 02 June 2006]
- Digital Preservation Testbed (June 2003) *From digital transience to digital durability; Preserving Emails*
<http://www.digitaleduurzaamheid.nl/bibliotheek/docs/volatility-permanence-email-en.pdf> [Accessed 02 June 2006]

- Donath, Judith (2004) *Visualizing E-mail Archives – draft*
<http://smg.media.mit.edu/papers/Donath/EmailArchives.draft.pdf> [Accessed 02 June 2006]
- Friedberg, Errol C; Hagler, H.K & Land, K.J. (June 2003) *How e-mail raises the spectre of a digital Dark Age* Nature, Vol 423, 19 June 2003
<http://www.nature.com/nature> [Accessed 02 June 2006]
- Greenspan, Robyn (September 2004) *Workplace IM Showing Growth* ClickZ Stats Professional
<http://www.clickz.com/stats/sectors/professional/article.php/3402961> [Accessed 02 June 2006]
- Harvey, Ross (forthcoming) *DCC Manual Instalment: Appraisal and Selection* (DCC)
<http://www.dcc.ac.uk/resource/curation-manual/chapters/appraisal-and-selection/>
- Hillesund, Terje (August 2002) *Many Outputs — Many Inputs: XML for Publishers and E-book Designers* Journal of Digital Information, Volume 3 Issue 1 Article No. 101
<http://jodi.tamu.edu/Articles/v03/i01/Hillesund/> [Accessed 02 June 2006]
- Iannella, Renato (March 1997) *The Resource Discovery Project* Ariadne, Issue 8
<http://www.ariadne.ac.uk/issue8/resource-discovery/> [Accessed 02 June 2006]
- ICA (1997) *The ICA Guide for Managing Electronic Records from an Archival Perspective* ICA
<http://www.ica.org/biblio.php?pdocid=3> [Accessed 02 June 2006]
- JISC (October 2002) *Freedom of Information Act 2000 : Snr Mgt : strategy and policy* JISC
http://www.jisc.ac.uk/index.cfm?name=pub_ibsm_foi [Accessed 02 June 2006]
- JISC (July 2001) *The Regulation of Investigatory Powers (RIP) Act 2000: E-mail and Telephone Monitoring* JISC Senior Management Briefing Paper 14
http://www.jisc.ac.uk/index.cfm?name=pub_smbp_ripa [Accessed 02 June 2006]
- Kelly, Brian (October 2005) *Web Focus: Must E-mail Die?* Ariadne, Issue 45
<http://www.ariadne.ac.uk/issue45/web-focus/> [Accessed 02 June 2006]
- Klimt, Brian & Yang, Yiming (2004) *The Enron Corpus: A New Dataset for E-mail Classification Research* ECML/PKDD 2004
<http://nyc.lti.cs.cmu.edu/yiming/Publications/klimt-ecml04.pdf> [Accessed 02 June 2006]
- Klimt, Bryan & Yang, Yiming (2004) *Introducing the Enron Corpus* First Conference on E-mail and Anti-Spam (CEAS) 2004 Proceedings
<http://www.ceas.cc/papers-2004/168.pdf> [Accessed 02 June 2006]
- Lavoie, Brian (January 2004) *The Open Archival Information System Reference Model: Introductory Guide* DPC Technology Watch Report
http://www.dpconline.org/docs/lavoie_OAIS.pdf [Accessed 02 June 2006]
- Lettice, John (December 2005) *Undelete those deleted emails, FOIA ruling tells Government* The Register
http://www.theregister.co.uk/2005/12/22/foia_undelte_ruling/ [Accessed 02 June 2006]
- Lord, Philip & MacDonald, Alison (2003), *e-Science Curation Report. Data curation for e-Science in the UK: an audit to establish requirements for future curation and provision*
http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf [Accessed 02 June 2006]
- Lowenthal, David (Summer 2005) *Stewarding the future* CRM: The Journal of Heritage Stewardship Vol 2 No 2 pp20-39
http://www.cr.nps.gov/crdi/publications/CRM_Vol2_02_Viewpoint.pdf [Accessed 02 June 2006]
- Lukesh, Susan S (1999) *E-mail and the Potential Loss to Future Archives and Scholars or The Dog That Didn't Bark* First Monday, Vol 4 No 9
http://www.firstmonday.dk/issues/issue4_9/lukesh/index.html [Accessed 02 June 2006]
- Moore, Regan et al (June 1999) *Collection-Based Long-Term Preservation*

<http://www.sdsc.edu/NARA/Publications/nara.pdf>
[Accessed 02 June 2006]

Norris, M. (October 2003) *Generic Policy for E-mail Retention and Disposal*
<http://www.lboro.ac.uk/computing/irm/generic-policy.html> [Accessed 02 June 2006]

Norris, Michael (November 2003) *Institutional Records Management & E-mail: Final Report*
<http://www.lboro.ac.uk/computing/irm/final-report.html> [Accessed 02 June 2006]

Perer, Adam; Shneiderman, Ben; & Oard, Douglass W. (2005) *Using Rhythms of Relationships to Understand e-mail Archives*
<http://hcil.cs.umd.edu/trs/2005-08/2005-08.pdf>
[Accessed 02 June 2006]

RLG (May 2002) *Trusted Digital Repositories: Attributes and Responsibilities*
<http://www.rlg.org/legacy/longterm/repositories.pdf>
[Accessed 02 June 2006]

Ross, Seamus (forthcoming) *Approaching Digital Preservation Holistically* in M Moss (ed.), *Information Management and Preservation*, Oxford, Chandos Press

Ross, S & McHugh, A (October 2005) *Audit and Certification of Digital Repositories: Creating a Mandate for the Digital Curation Centre (DCC)*
http://www.rlg.org/en/page.php?Page_ID=20793#article1 [Accessed 02 June 2006]

Rothenberg, Jeff (January 1999) *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation* CLIR, pub 77
<http://www.clir.org/pubs/abstract/pub77.html>
[Accessed 02 June 2006]

Shepard, Thom (December 1998) *Universal Preservation Format (UPF): Conceptual Framework* RLG Diginews Volume 2, No.6
<http://www.rlg.org/preserv/diginews/diginews2-6.html#upf> [Accessed 02 June 2006]

Sherriff, Lucy (July 2005) *e-mail? That's for old farts!* The Register

http://www.theregister.co.uk/2005/07/28/email_for_old_farts/ [Accessed 02 June 2006]

Squeo, Anne Marie (December 2005) *Oh, Has Uncle Sam Got Mail* in The Wall Street Journal Online 29 December 2005
http://online.wsj.com/public/article/SB113581938626033499-xNP7F7iqAatGMjivCNMuy6GOH2M_20061229.html?mod=blogs [Accessed 02 June 2006]

The National Archives (2002) *Functional Requirements for Electronic Records Management Systems*
<http://www.nationalarchives.gov.uk/electronicrecords/reqs2002/>. [Accessed 02 June 2006]

Thormeyer, Rob (February 2006) *NARA finalizes rule on short-term e-records* in Government Computer News Feb 2nd 2006
http://www.gcn.com/vol1_no1/daily-updates/38316-1.html [Accessed 02 June 2006]

Tyler, Joshua R; Wilkinson, D. M.; Huberman, B. A. (2003) *E-mail as Spectroscopy: Automated Discovery of Community Structure within Organisations* HP Laboratories
<http://www.hpl.hp.com/research/idl/papers/email/email.pdf> [Accessed 02 June 2006]

U.S .Department of Defense Standard 5015.2 'Design Criteria Standard For Electronic Records Management Software Applications' (2002)
<http://jitic.fhu.disa.mil/recmgt/p50152s2.pdf>
[Accessed 02 June 2006]

Viégas et al (2004) *Digital Artifacts for Remembering and Storytelling: PostHistory and Social Network Fragments* Proceedings of the 37th Hawaii International Conference on System Sciences
http://alumni.media.mit.edu/~fviegas/papers/posthistory_snf.pdf [Accessed 02 June 2006]

Walsh, Norman (September 2002) *XML: One Input — Many Outputs: a response to Hillesund* in Journal of Digital Information, Volume 3 Issue 1 Article No. 165, 2002-09-12
<http://jodi.tamu.edu/Articles/v03/i01/Walsh/>
[Accessed 02 June 2006]

Wheatley, Paul (October 2001) *Migration - a CAMiLEON discussion paper* Ariadne, Issue 29
<http://www.ariadne.ac.uk/issue29/camileon/>
[Accessed 02 June 2006]

Wall Street firms fined \$8.25 million for deleting e-mail (December 2002) OUT-LAW News, 04/12/2002
<http://www.out-law.com/page-3168> [Accessed 02 June 2006]

Firms face fines for discarding e-mail traffic (August 2002) OUT-LAW News, 02/08/2002
<http://www.out-law.com/page-2820> [Accessed 02 June 2006]

JP Morgan fined \$2.1 million over e-mail record keeping (February 2005) OUT-LAW News, 15/02/2005
<http://www.out-law.com/page-5303> [Accessed 02 June 2006]

Morgan Stanley offers \$15 million to make up for missing emails (February 2006) OUT-LAW News, 21/02/2006
<http://www.out-law.com/page-6656> [Accessed 02 June 2006]

Freedom of Information Act 2000 The Stationery Office Limited
<http://www.opsi.gov.uk/acts/acts2000/20000036.htm> [Accessed 02 June 2006]

DCC Information Day Report - University of Ulster/Queens University Belfast (December 2005) DCC
http://www.dcc.ac.uk/events//info-day-2005-december/belfast_info_day.pdf [Accessed 02 June 2006]

ERPA-Study: Council of Europe (2003) ERPANET
http://www.erpanet.org/studies/docs/erpaStudy_COE.pdf
[Accessed 02 June 2006]

ERPA-Study: Theater Instituut Nederland (2004) ERPANET
http://www.erpanet.org/studies/docs/erpastudy_TIN.pdf [Accessed 02 June 2006]

MoREQ Model Requirements for the Management

of Electronic Records (March 2001) Office for Official Publications of the European Communities as INSAR Supplement VI, ISBN 92-894-1290-9 also available at
<http://cornw.co.uk/moreqdocs/moreq.pdf> [Accessed 02 June 2006]

New Digital Archive at The National Archives (undated) TNA
http://www.nationalarchives.gov.uk/preservation/digitalarchive/pdf/project_background.pdf [Accessed 02 June 2006]

Why Instant Messaging Management? (May 2004) The ePolicy Institute
<http://www.epolicyinstitute.com/imr/intro.pdf>
[Accessed 02 June 2006]

Data Protection Act 1998 The Stationery Office Limited
<http://www.opsi.gov.uk/acts/acts1998/19980029.htm> [Accessed 02 June 2006]

E-mail Policies and Practices: An Industry Study Conducted by AIIM International and Kahn Consulting, Inc. - Managing Email in the New business Reality (2003).
<http://www.aiimhost.com/membership/EmailSurvey1-9.pdf> [Accessed 02 June 2006]

Human Rights Act 1998 The Stationery Office Limited
<http://www.opsi.gov.uk/acts/acts1998/19980042.htm> [Accessed 02 June 2006]

ISO 14721:2003 Open Archival Information System (OAIS) Reference Model (February 2003) ISO
<http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=24683&ICS1=49&ICS2=140&ICS3> (purchase) [Accessed 02 June 2006]

Regulation of Investigatory Powers Act 2000 The Stationery Office Limited
<http://www.opsi.gov.uk/acts/acts2000/20000023.htm> [Accessed 02 June 2006]

E-mail Archiving: A Legal Perspective, (July 2003) Team Discovery Ltd
<http://www.teamdiscovery.co.uk/cgi-bin/newslst.pl?bid=41&num=4> [Accessed 02 June 2006]

The Telecommunications (Lawful Business Practice) (Interception of Communications) Regulations (2000) The Stationery Office Limited
<http://www.opsi.gov.uk/si/si2000/20002699.htm>
[Accessed 02 June 2006]

Freedom of Information (Scotland) Act (2002) The Stationery Office Limited
<http://www.itspublicknowledge.info/legislation/act/foiactcontents.htm> [Accessed 02 June 2006]

Convention for the Protection of Human Rights (2003) Council of Europe

<http://www.echr.coe.int/NR/rdonlyres/D5CC24A7-DC13-4318-B457-5C9014916D7A/0/EnglishAnglais.pdf> [Accessed 02 June 2006]

DIRECTIVE 1999/93/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 13 December 1999 on a Community framework for electronic signatures (December 1999) Official Journal of the European Communities 19. 1. 2000
http://europa.eu.int/eur-lex/pri/en/oj/dat/2000/l_013/l_01320000119en00120020.pdf [Accessed 02 June 2006]

Key external resources

Collection Based Long Term Preservation

<http://www.sdsc.edu/NARA/>

Digital Preservation at the National Archives of Australia

<http://www.naa.gov.au/recordkeeping/preservation/digital/applications.html>

Digital Preservation Coalition (DPC)

<http://www.dpconline.org/>

Digital Preservation Testbed

<http://www.digitaleduurzaamheid.nl/>

ENRON e-mail dataset

<http://www.cs.cmu.edu/~enron/>

ERPANET

<http://www.erpanet.org/>

European Court of Human Rights

<http://www.echr.coe.int/echr/>

Expertise Centrum eDavid

<http://www.expertisecentrumdavid.be/>

Gmail

<http://mail.google.com/>

Information Commissioner's Office

<http://www.ico.gov.uk/>

Institutional Records Management & e-mail project

<http://www.lboro.ac.uk/computing/irm/>

Office of Public Sector Information

<http://www.opsi.gov.uk/>

PADI:emails

<http://www.nla.gov.au/padi/topics/47.html>

Public Citizen

<http://www.citizen.org/litigation/>

RFC Archive

http://www.ietf.org/iesg/1rfc_index.txt

RLG Diginews

<http://www.rlg.org/preserv/diginews/>

Scottish Information Commissioner

<http://www.itpublicknowledge.info/>

Sourceforge

<http://sourceforge.net/>

Appendix 1 – Sample institutional guidelines, advice, and policies on managing e-mails

The implementation of the Data Protection Act in 2000, and particularly the Freedom of Information Act in January 2005, reinforced the need for guidelines on the use of e-mail in UK organisations. The majority now have at least basic guidelines in place regarding the use of e-mail and e-mail systems, and such guidelines, advice and policies on managing e-mails are an increasing aspect of records management activities or IT support. This section introduces a selection of work, guidelines, advice and policy on e-mail management originating from the UK.

Policy recommendations and outcomes of the Institutional Records Management and E-mail project

As part of their 2003 Supporting Institutional Records Management programme, JISC funded a six-month project at Loughborough University to examine the use of e-mail in HE and FE institutions: the Institutional Records Management and E-mail (IRME) project.⁷¹ In recognition of e-mail's growing importance as a unique communications and record-producing tool, the project examined e-mail and attachments as records and how they could be integrated into the University's wider records management system.

During the background research, project staff carried out a survey into e-mail policies at other UK HE institutions.⁷² The survey drew 39 responses that indicated the sector was in the early stages of resolving the issues presented by e-mail archiving. E-mail was the subject of a number of general policy documents concerned with the use and

management of e-mail accounts, but the survey was unable to identify any institution that had a well-defined and active policy in place specifically for e-mail archiving.

The project deliverables included a generic policy on e-mail retention which recognises e-mail is a form of record that the University has a responsibility to manage.⁷³ The main points of the retention policy concern the individual's responsibilities to identify and categorise e-mail records of value, apply relevant contextual information and metadata, and manage e-mail records in a manner that ensures their integrity. Similarly, the University has a responsibility to provide guidance on the categories of e-mail to be retained and on retention schedules, provide a viable archiving system, and provide training and support. The team concluded that:

It is essential to recognise that e-mail correspondence is an institutional record that should be managed under a records management system like any other record. The University should therefore develop an integrated corporate records management policy which includes all records of value irrespective of the medium on which they are held. Part of the policy should deal with electronic records and as a subset of this, e-mail records should be treated as potentially legally admissible documents and a policy developed appropriately.⁷⁴

The project stressed the need for e-mail policies, separate from the general Acceptable

⁷¹Institutional Records Management and E-mail project web site <http://www.lboro.ac.uk/computing/irm/>.

⁷²Final Report of the Institutional Records Management and E-mail project

<http://www.lboro.ac.uk/computing/irm/final-report.html>.

⁷³Op cit; see also the Institutional Records Management and E-mail project *Generic Policy for E-mail Retention and Disposal*

<http://www.lboro.ac.uk/computing/irm/generic-policy.html>

⁷⁴Final Report of the Institutional Records Management and e-mail project, op cit.

Use Policy, to provide guidance on acceptable use of e-mail and address the particular responsibilities that different staff members may hold. Furthermore, compulsory training should be provided that covers document appraisal and record schedules, as well as the use of an e-mail client and e-mail archive.

Policies, guidelines and advice from a sample of universities

Following on from the IRME project, it appears that almost every University now has a policy or guidelines that address e-mail management from some perspective. A sample examination of policies reveals that some are more detailed than others.

The University of Warwick has produced a brief e-mail guidance/policy note, *Electronic Mail Policy, v1.3.1*, that covers use of official e-mail accounts, access to personal inboxes during a period of absence, account closure when a user leaves the university, retention and archiving (in this instance, print to paper) and spam.⁷⁵ Warwick's *Email Best Practice* follows similar lines but provides slightly more detail on each of the topics covered.⁷⁶ The University of Edinburgh, on the other hand, covers the topic of *Managing your E-mails* in far greater depth, providing detailed advice on Data Protection and FoI, responsibilities for e-mail management, issues to consider in writing e-mails and sending attachments, retention or deletion, retention periods, storage locations and formats, encryption, absenteeism, spam, and use of the e-mail system for personal e-mails.⁷⁷ Aberdeen University provides

additional advice on managing e-mail threads, legal admissibility, privacy and security in its *Guidelines For The Management Of University E-Mail*.⁷⁸ The University of Bath e-mail policy originates from computing services and focuses mainly on acceptable use with links through to separate web pages on FoI, Data Protection, RIR, the Human Rights Act, and the Computer Misuse Act (1990).⁷⁹ The University of Aberystwyth's *Policy on the use of e-mail* covers general e-mail considerations and advises that although e-mail is an informal means of communication it is nonetheless a form of publication with legal liability. For long-term accessibility and record-keeping purposes, it recommends that e-mail should be transferred to another electronic environment or printed out to paper.⁸⁰

This sample overview of current practice in institutional policies for e-mail use and management appears fairly representative of institutional e-mail policy as a whole, in that policies vary wildly in terms of scope and contents. Practices appear to have developed since the IRME project survey, possibly informed by their findings. Some institutions (most notably Edinburgh) conform with the IRME advice, providing detailed advice on storing e-mails in suitable storage formats, and explicitly advocating the integration of electronic mail records into the overall record keeping infrastructure.

⁷⁵University of Warwick: *Electronic Mail Policy, v1.3.1* (June 2004)
http://www2.warwick.ac.uk/services/its/helpfaq/policies/e-mail_guidance.pdf.

⁷⁶University of Warwick: *Email Best Practice* (April 2006)
<http://www2.warwick.ac.uk/services/its/facilities/e-mail/bestpractice/>.

⁷⁷University of Edinburgh (UoE): *Managing Your Emails v9*(Jan 2005)
<http://www.recordsmanagement.ed.ac.uk/InfoStaff/RMst>

[aff/ManagingEmail/ManagingEmailMainV9.pdf](#). The approach taken at UoE was the subject of a presentation by Susan Graham at the DCC workshop on Curating Emails held in Newcastle-Upon-Tyne in April 2006 and is available from <http://www.dcc.ac.uk/events/ec-2006/>.

⁷⁸University of Aberdeen: *Guidelines For The Management Of University E-Mail* (Sept 2005)
<http://www.abdn.ac.uk/central/records-management/e-mail.shtml>.

⁷⁹University of Bath: *Computing Services E-mail Policy* (Feb 2003) <http://www.bath.ac.uk/bucs/policies/e-mail.shtml>.

⁸⁰University of Wales, Aberystwyth: *Policy on the use of E-mail* (March 2004)
<http://www.aber.ac.uk/infopolicies/e-mail-policy.shtml>.

Guidelines on developing an e-mail policy from UK The National Archives

The National Archives has issued guidelines for government agencies to follow in *Developing a Policy for Managing e-mail*.⁸¹ The guidelines covers two main areas that an organisation should consider in developing an e-mail policy: appropriate use of e-mail, and; how e-mail should be managed within an organisation. The purpose and benefits of such a policy are clearly stated:

“Producing an e-mail policy will clarify an organisation’s position on how e-mail should be treated within the organisation. This type of clarification is necessary to help ensure that e-mail is used in a way that an organisation abides by its legislative requirements, maintains e-mail records relating to business and encourages staff members to write e-mail messages that do not confuse the recipient”.

The guidelines encourage organisations to provide advice on when to use e-mail, on managing e-mail messages, managing public and shared mailboxes, identifying and managing records, writing message content, and naming conventions. Embedding these institutional requirements in a policy helps ensure that e-mail is afforded the same status as written documents or digital records in another form, and helps combat user perception of e-mail as an informal and personal tool.

⁸¹*Developing a Policy for Managing E-mail* (2004) available from The National Archives web site http://www.nationalarchives.gov.uk/electronicrecords/advance/pdf/managing_emails.pdf

Appendix 2 - E-mail Curation and Preservation In Action

Practical activities in e-mail curation and preservation vary considerably. Although a growing number of institutions are developing policies and guidelines to address the proper use and creation of e-mails from both an individual and an organisational perspective (particularly in the light of new legislation and recent and financially-damaging 'e-mail scandals' referred to in the introduction), very few have established proper e-mail archives that fully address the principles of curation and preservation and even fewer e-mail archives are publicly available. A small number of technical solutions to preserve e-mails for the long-term have been developed and are publicly available, but they are yet to be widely implemented.

This section introduces some potential technical approaches to facilitating or ensuring the long-term preservation of e-mail messages and offers a brief case study on the re-use of e-mails in the so-called ENRON corpus.

A2.1 Technological solutions for e-mail preservation

Growing recognition of e-mail as a record type and the problems caused by proprietary e-mail formats have led to the research and development of strategies and tools specifically for preserving or archiving e-mails.

Dutch National Archives – Testbed XMaiL

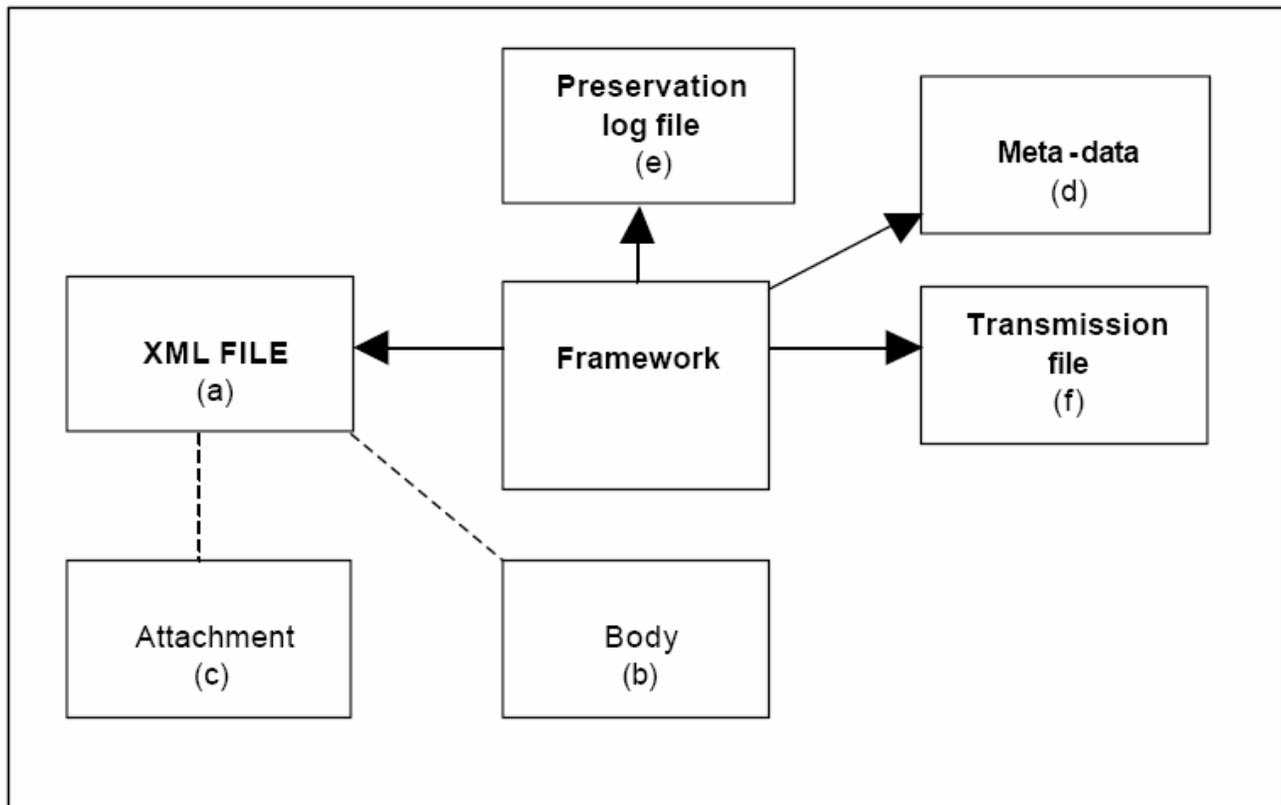
The Dutch government project Testbed Digitale Bewaring/Digital Preservation Testbed was part of a larger Digital Longevity programme that sought to investigate several aspects of digital longevity for government agencies.⁸² The Testbed project ran from 2000 to 2003 and carried out a series of experiments to investigate the effects of different

preservation strategies on different types of digital records commonly found in the office environment. The project also produced recommendations on functional requirements for a preservation system, a decision model to help select a suitable preservation approach for different types of records, authenticity requirements for different record types, and costing information.

Based on the outcome of the experiments, the project team published a series of recommendations entitled '*From digital transience to digital durability*', one for each of the record types investigated. One of the record types was e-mail. The recommendations discussed and provided advice for records creators, managers, record-keepers and IT staff on their roles in the curation and preservation processes. The team concluded that the best way to preserve e-mails for the long-term was to use an XML based solution as part of a wider encapsulation approach that linked multiple representations of the message with appropriate metadata and a 'preservation log book' together in a so-called 'e-mail preservation object'.

⁸²Digital Preservation Testbed web site
<http://www.digitaleduurzaamheid.nl/home.cfm>.

Figure 1: Diagrammatic representation of the Testbed e-mail preservation object



⁸³The e-mail preservation object reproduced above comprises at least four components linked together by a central Framework (or linkage stem). Solid lines indicate compulsory components and broken lines indicate optional ones. The XML file (a) may be a single file containing all of the required data, or may contain only header information and link to two separate XML files containing the message content (b) and any decoded attachments (c). Further files contain metadata (d), a preservation log – effectively an audit trail of actions carried out on the record object (e) – and a copy of the original RFC 2822 transmission file (f).

The project team also developed software to mark-up new and existing e-mail messages in

XML. The *XMaiL* installation package is available for free download from the Testbed project website.⁸⁴ It is open source with a BSD license that allows commercial use of the software as well as redistribution and use of the binary and source forms. The plug-in cannot be immediately installed on users desktops and requires configuration by IT staff. It is designed for use with Microsoft Outlook, although the concept behind the plug-in can be applied to any type of e-mail client. XML message files generated by the plug-in conform to an XML schema. Additional contextual metadata is included in the XML file, which also allows for automated record-keeping filing based on the categories selected. Additional information about the sender and recipient(s) is

⁸³© Digital Preservation Testbed, The Hague, 2003. *Digital Preservation Testbed: From digital volatility to digital permanence. Preserving E-mail* (2003) op cit.

⁸⁴XMaiL software: <http://www.digitaleduurzaamheid.nl/index.cfm?paginakeuze=299>.

included in the file by linkage and copying of the data in the compulsory address book entries in the Outlook system.

A technical description of the *XMail* software and a flash animation illustrating how the software works are also available from the project web site, as is the XML Schema.⁸⁵

Antwerp City Archives – E-mail Preservation Template

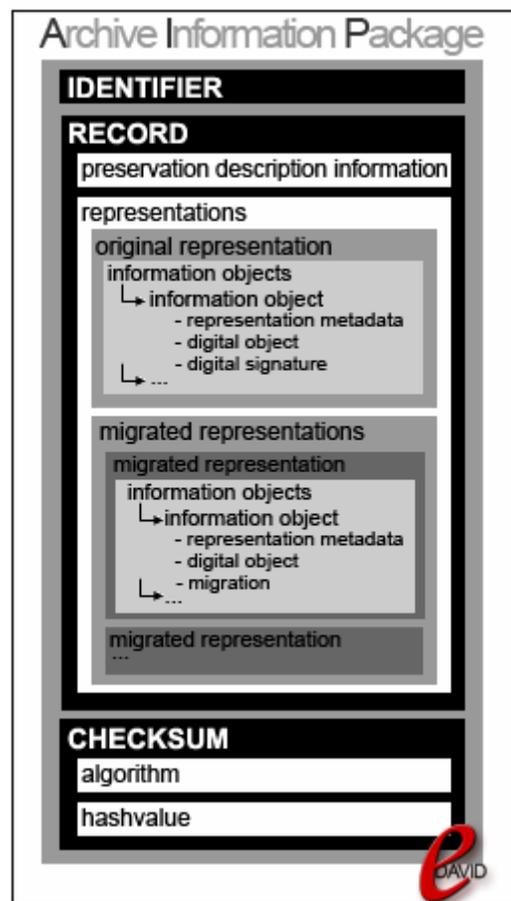
Antwerp City Archives, together with the Interdisciplinary Centre for Law and Informatics of the K.U.Leuven in Belgium, carried out a four year project from 1999 to 2003 to research digital durability in a government environment. The DAVID project – Digitale Archiveren In Vlaamse Instellingen en Diensten – lasted for four years and produced a series of guidelines and reports on digital archiving and digital record keeping from both archival and legal perspectives.⁸⁷ Of particular interest to the project were web sites, e-mails, and the electoral register.

Expertise Centrum eDavid carries on the work started by the DAVID project. The original project report '*Archiving E-mail*' from 2002 was updated by eDavid in 2006 in the report '*Filing and Archiving E-mail*'. This report considers in great depth the European and Belgian legal and archival issues that affect the infrastructure of an e-mail management and archiving approach. A Best Practice for e-mail archiving is proposed that draws from these discussions and comprises selection, registration, classification and storage, possible migration to XML and integration with the overall organisational record-keeping infrastructure. The approach has been tested in a pilot project that used a plug-in similar to that

develop by the Testbed project. An e-mail-XML DTD and Schema, and a related XSL Stylesheet have also been developed; these files are designed to work in tandem with the DAVID macro plug-in for customising Microsoft Outlook and producing an XML representation of the messages.⁸⁸ As with the Testbed system, the plug-in adapts the user interface for incoming and outgoing emails, enables additional (but minimal) metadata collection, and assists in storing e-mails offline in the correct location.

Focusing on the two pillars of authenticity and digital repositories, the eDavid project has developed a storage architecture for digital records based around the OAIS AIP model.

Figure 2: Structure of an eDAVID AIP⁸⁶



⁸⁵Technical description, Flash animation and Schema available from <http://www.digitaleduurzaamheid.nl/index.cfm?pagina&uze=185>.

⁸⁶© eDavid/expertisecentrum DAVID, 2005.

⁸⁷DAVID & eDavid project website <http://www.expertisecentrumdavid.be/>.

⁸⁸http://www.expertisecentrumdavid.be/davidproject/downloads/DAVID_emailsjabloon.zip.

This form of encapsulation again utilises XML.

The AIP consists of three main sections: a persistent identifier, the record, and checksum information. It conforms to the OAIS template for information packages and contains the Preservation Description Information (PDI) necessary to manage the object over time. Multiple representations of the record can be stored in the record object section, accompanied by representation information and information about any digital signatures. Finally, the AIP contains a checksum algorithm and hash value that can be used to establish the fixity and validity of the package as necessary.⁸⁹

National Archives of Australia – XENA software suite

The preservation of e-mail records is a component of the Digital Preservation Project at the National Archives of Australia (NAA). The National Archives has developed an integrated approach to the storage and preservation of archival records, including normalisation and transfer of records into a digital repository, using a suite of Open Source tools: *XENA*, *DPR*, and *QUEST*.⁹⁰

XENA - XML Electronic Normalisation of Archives - converts digital records into XML formats the NAA has developed or selected as most suitable for long-term preservation.⁹¹ XML formats have been developed for a range of source formats, including emails in *.pst, *.trim, and *.mbox formats. Different XML is produced depending on the content type of the message, whether that is plain text or HTML, and attachments are treated according to their

original file type. A bit stream version of the file is also saved. Both are wrapped in metadata. The bit stream version is the most complete version of the submitted record but relies upon the original computing environment for accurate rendition. The XML version is less accurate but its performance is considered as close to the original as currently possible.

The *DPR* – Digital Preservation Recorder – captures an audit trail of the processes the digital object undergoes during the preservation process and enables the NAA to record a complete life history of the objects in its care. *QUEST* – QUery Electronic STorage - creates and maintains links between the objects stored in the repository and the metadata that accompanies them.⁹² It also retrieves Archival Information Packages from the digital repository. Together, these three tools and the digital repository enable the Archives to manage and preserve records transferred to them, integrating e-mail preservation with that of other record types.

San Diego Supercomputer Centre (SDSC)

In 1999, the SDSC carried out the Collection-Based Long Term Preservation research project to develop and test approaches for preserving the organisation of digital collections simultaneously with the digital objects that comprise the collection.⁹³ Sponsored by NARA, the project developed a solution to cater for the explosion in various types of electronic records created by US Government agencies. A collection of e-mail postings from the Usenet groups at SDSC was used to demonstrate the applicability and scalability of the approach for e-mail messages. The e-mail collection was comprised of a million messages without attachments, with a raw size of 2.52 GB. These messages were RFC 1036 standard

⁸⁹For a complete review of the approach, see Bousdrez, F. *Digital containers for shipment into the future* (2005) http://www.expertisecentrumdavid.be/docs/digital_containers.pdf.

⁹⁰For information about XENA, DPR and QUEST, see <http://www.naa.gov.au/recordkeeping/preservation/digital/applications.html>.

⁹¹Available for download from <http://xena.sourceforge.net/download.html>.

⁹²QUEST has not yet been released to the public. The expected URL is <http://quest-archiv.sourceforge.net/>.

⁹³Moore, R. et al, *Collection-Based Long term Preservation* (1999) <http://www.sdsc.edu/NARA/Publications/nara.pdf>.

messages, RFC 1036 being a standard for the Interchange of network news messages among USENET users, and not the RFC 2822 format to which true e-mail messages conform.^{94 95}

XML images of the records were made that complied with a defined e-mail DTD. Unique tags were added to the beginning and end of each record, and the e-mail message bodies were concatenated into twenty-five message body data files, one for every 40,000 messages. These were stored in the Persistent Archive, which uses commercially available storage systems – most notably IBM's High Performance Storage System as the Archival Storage System - augmented by application level software developed by the SDSC. This includes the *Storage Resource Broker (SRB)*, which writes the containers from the cache system where they are initially held into the Archive. A web-based interface then allows access to the messages and other objects in the Archive.

The software and approach developed in this project is still under development and is not yet available to external parties.

Commercial software

Several commercial software solutions claim to archive and preserve e-mail. It is difficult to establish exactly how reliable these systems may be without undertaking a thorough examination of their infrastructure and functionality. Most software solutions are developed to work on an existing collection of messages from an organisational level and have arisen from the high-profile cases involving loss or retention of inappropriate messages by

companies, particularly in the US. Few solutions have been developed to operate on a smaller and personal scale, as required by individuals wishing to ensure their emails are not locked into a proprietary format. One such solution is the '*Emailchemy*', commercial software that converts messages from 19 different e-mail applications (including Mozilla, Outlook express, Netscape, Mulberry and Eudora) into RFC 2822 or CSV files. It can write to folders with individual RFC 2822 files (.txt or .eml) or to RFC 2822 mailboxes ("mbox" format or "UNIX style").⁹⁶ Further action is then required to turn the mbox files into individual messages and to implement persistent storage, search, and access facilities. The effect of this software on the archival authenticity and integrity of the emails has not been publicly tested.

A2.2 E-mail re-use in publicly available collections

Very few e-mail archives or collections of e-mail messages are yet publicly available. A notable exception to this is the ENRON corpus, obtained by investigators during the ENRON accounting fraud scandal uncovered in 2001. The Federal Energy Regulatory Commission (FERC) was charged with investigating ENRON and initiated a review of all available ENRON data, including e-mails, to determine the extent of ENRON's misdeeds. The collection of e-mails, known as the ENRON corpus, has since been made publicly available in two forms: the initial 'FERC' set of data, and a 'cleaned' set.

The FERC set contains the data released to FERC during their investigation.⁹⁷ It is searchable and contains 92% of ENRON's staff

⁹⁴RFC 1036 Standard for Interchange of USENET messages: <http://www.faqs.org/rfcs/rfc1036.html>.

⁹⁵USENET is a worldwide distributed discussion system consisting of a set of predefined 'newsgroups' – subjects – to which users post messages. The main difference between e-mail and USENET is that e-mail can be used for both public and private correspondence, whereas USENET messages are public and can be read by anyone with access to the USENET system.

⁹⁶Available from <http://www.weirdkid.com/products/emailchemy/index.html>.

⁹⁷Available from <http://fercic.aspensys.com/members/manager.asp>. (Username and password required)

e-mails. Data is represented as plain text, although OCR images of e-mails with a print-like appearance can be generated on-demand. Attachments are also available. A 'cleaned' set of the data was later released that does not include attachments but there are issues surrounding this effort that mean the collection and its contents may be misinterpreted.⁹⁸ Some messages were deleted 'as part of a redaction effort due to requests from employees and valuable data contained in the initial set that identifies, for example, BCC recipients in 'sent' messages, has been destroyed. Such information can indicate hierarchies, conspiracy, and unwarranted, unnecessary, or illegal involvement in official business. Attachments have been removed and recipients names, when they appeared in the header fields, were replaced by valid e-mail addresses along the lines of firstname.lastname@enron.com. These actions were originally carried out to address 'integrity problems' in the dataset, but they have resulted in damage to the archival integrity of the set. Re-usability of the archive is not straightforward, as the *.fdb files that enable access via an inbox mechanism have been removed and prevent easy identification of which messages were 'replied to'. Finally, 'records' in the dataset are not clearly identified and cannot easily be distinguished from non-records, i.e. messages to and from listservs, personal messages to family and friends, and inappropriate material, without accessing the message content.

The ENRON corpus is available 'as a resource for researchers who are interested in improving current e-mail tools, or understanding how e-mail is currently used'.⁹⁹ It has been used as source material for a number of e-mail visualisation experiments, natural language processing investigations, and research examining methods for automatic

categorisation of e-mail into folders. Although the value of the original dataset is not disputed, use of the cleansed set for research into e-mail usage may, for the reasons specified above, produce flawed results and possibly even skew the direction of future research as the dataset is not a true representation of an e-mail collection but is one of the only ones currently available for testing.

⁹⁸Available from <http://www.cs.cmu.edu/~enron/>.

⁹⁹Ibid