Clipping Mercury's Wings: The Challenge of Email Archiving

Mark Brogan

Dr Mark Brogan teaches and researches information technology and information science in the School of Computer and Information Science at Edith Cowan University. He has published previously on the theory and practice of relational database archiving and recordkeeping practices in Australia's Internet service provider industry.

This article investigates the changing landscape of email archiving. It explores the gap between user behaviours, professional frames of good practice email management and various systems solutions.

The paper argues for alignment of methods, procedures and business solutions with ethnographic and survey evidence of how users respond to the challenge of email management in circumstances of 'information overload.' The term 'information overload' emerged in connection with computer-mediated communication in the 1990s. A significant research literature has since grown up around the causes and implications of information overload, including email induced information overload. The author argues that evidence-based interpretations of end-user email filing behaviour sit uncomfortably with recordkeeping approaches to email management based on 'recordness'. The author concludes that approaches to email archiving based on volition and 'recordness' may need to be abandoned in favour of approaches that make sense in terms of what is known about user behaviour.

Introduction

Recent case studies in digital recordkeeping have focused attention on email archiving and managing messages as records. In *Citizens for Responsibility and Ethics in Washington v. Executive Office of the President, et al.*, a non-government public interest advocacy claimed White House violation of the *Presidential Records Act.*[1] According to the plaintiffs, White House staff failed to archive White House emails sent and received between 2003 and 2008. Elsewhere in the United States, the Governor of North Carolina announced a review of the use of state-owned email systems amid claims that by allowing uncontrolled deletion of emails and other electronic text communications on BlackBerry handheld units, his administration had violated state records law.[2]

In a recent Australian case study, the 'Fong–Burke Email Affair', digital forensic methods were used by an anti-corruption 'watchdog', the Corruption and Crime Commission (CCC), to recover contentious email sent to a political lobbyist by the Director General of the Western Australian Health Department, Dr Neale Fong.[3] The existence of these emails had been the subject of denials by Dr Fong. Forensic investigation revealed thirty-three email contacts, some involving Health Department business. In common with the North Carolina case, Dr Fong's use of a BlackBerry mobile agent for the management of email became an important focus of investigation.[4]

Whilst cases involving personal digital assistants (PDAs) suggest the 'cutting edge' nature of archival email management, there is long pedigree to this issue in public policy and archival thinking. Beginning with the PROFS affair[5] in the 1980s, email and message management in public and private sector corporations has provided the basis of lively debate within the records, archives and information systems communities. In an article published in *Archives and Manuscripts* in 1994, American theorist on archives David Bearman laid out the accountability significance of email. Bearman introduced the problem with the following observation:

> The question of how to manage electronic mail as a record is one that will confront management of every contemporary organization within the next few years. The impetus may be to document what the organization has done to make better decisions, enforce contracts or to avoid claims, or it may be to reduce risks by destroying electronic records as

> soon as they are not required for operational reasons. In either case, we require a framework to help us ask the question of how to assure that electronic mail results in the creation of a record and how to manage records created by electronic mail communications systems over time.[6]

Bearman argued that tactics were needed to guide organisational responses to the problem of email management. More than a decade on, email and digital messaging more generally remain a vexed area of recordkeeping. In a prologue to its revised guidelines for selection of organisational retention policy, the Sedona Conference noted both the importance and challenge posed by email management:

> Electronic mail ('Email') is of vital importance to the productive efforts of an enterprise and its use is growing exponentially. In 2005, the average user processed 75 emails a day and the Radicata Group estimates that corporate email traffic per user has increased at a rate of 33% per year since then. Projections are that worldwide traffic in 2006 was at the rate of 183 billion messages per day.[7]

According to a recent UK Information Management industry survey that tested confidence in email archiving, two-thirds of managers 'have little or no confidence' that emails related to business decisions and obligations are 'recorded, complete and recoverable.'[8] The same survey indicated that public sector managers were less confident than their private sector counterparts.

Framing email archives management

From mobile technologies that do not readily integrate with records and information management systems, to the more familiar territory of user behaviour and its consequences for the record, email management throws up challenges to archival methods. These challenges begin with the absence of a shared understanding between stakeholders about the nature and purpose of the archival enterprise and warrants for email archiving. Information technology professionals, users and the recordkeeping community frame the problem of email archiving differently, giving rise to competing notions of 'best practice' in email archives management.

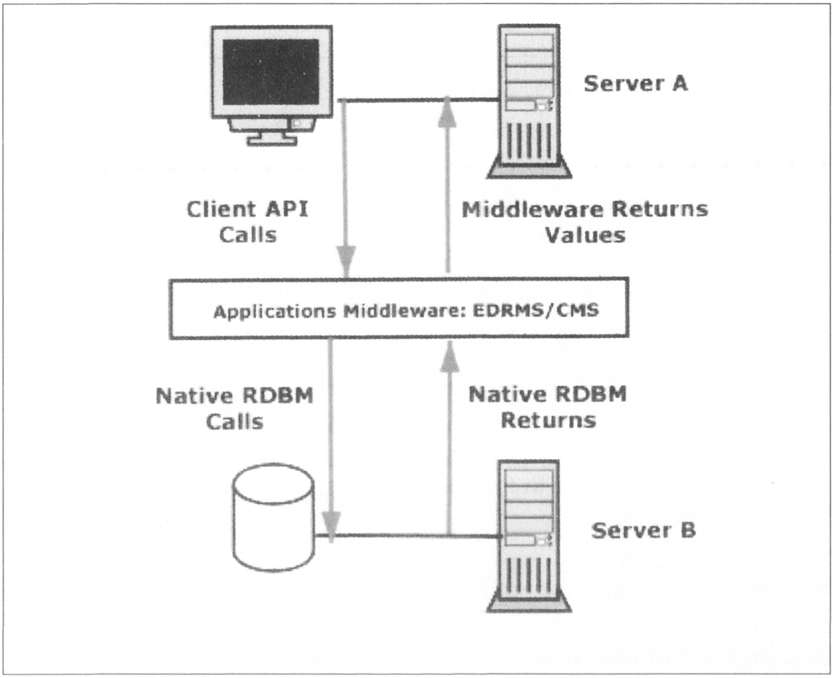

Figure 1. Information Architecture: Middleware

Shaped by a notion of the record as a 'document plus authenticating processes', including intent on the part of the document creator, many recordkeeping professionals adhere to a notion of 'best practice' that relies on user filing of record mails to online corporate stores via applications middleware. Within the store, retention and disposal policy operating on folders and files ensures the retention of record and archival mail. Figure 1 shows the information architecture of a middleware application of this kind found in an enterprise electronic document and records management systems (EDRMS) or content management system (CMS).

Alternatively, retention and disposal policy might be applied to user folders organised around an enterprise's business classification scheme, enabling archival copy to be captured into a content management system or replicated on an email server each time a user files a mail to a client folder with archival retention status. In 2007, Microsoft introduced this capability to its email client MS Outlook via the Microsoft Office SharePoint Server (MOSS) integration:

> In order to help users to declare email messages as records using the familiar Office Outlook 2007 client software, Office SharePoint Server 2007 and Microsoft Exchange Server 2007 have been tightly integrated. IT departments can create organizational folders in Exchange Server 2007 that map to business functions, which can be pushed out to a User's Outlook 2007 client using Group Policy. Users can simply drag and drop email messages into these folders from their client computer, causing Exchange Server 2007 to auto-copy these messages to Office SharePoint Server 2007.[9]

In another, somewhat anachronistic, but still encountered view of best practice email management, record emails are printed to hard copy and filed into paper-based filing systems. These systems form the locus of recordkeeping in a hybrid information and records management operating environment.

End-user computing behaviour and email management

In their own way, each of these assertions of best practice is problematic if underlying assumptions about user email management behaviour are found to be unsupported by evidence-based research. A number of studies have sought evidence-based explanations of user behaviour and sought to explain the implications of such behaviour for good practice email management.[10] Most originate from the field of end-user computing.

In an early ethnographic study of email management behaviour aimed at investigating the problem of information overload, Whittaker and Sidner investigated a stratified sample of users in Lotus Development Corporation. They found that issues with filing, task management and asynchronous communication (for example, email conversation threading) were typical of users. An assumption implicit in many good practice email policy and procedural guidelines, namely, that end users would routinely and reliably file email, was confounded by survey findings. Users were found to be uncertain about how to file email, resulting in most email being in a 'holding pattern' while users figured out what to do with it. Where users created folders, 35% were found by Whittaker and Sidner to contain only one or two items:

> Not only do these tiny 'failed folders' not significantly reduce the complexity of the inbox, the user has the dual overheads of (a) creating them in the first place, and (b) remembering multiple definitions every time there is a decision about filing a new inbox item.[11]

The authors concluded that return on investment in creating folders

> may not be great: folders can be too large, too small or they may be too numerous for people to remember their individual definitions. As a consequence, folders may be of little use either for retrieval or viewing related messages together.[12]

Whittaker and Sidner identified various strategies employed by users for dealing with information overload in email. They expressed these strategies in terms of behavioural types: *no filers, spring cleaners* and *frequent filers. No filers* were email users who relied upon full text searching of the inbox to find mails and typically did not file mail received; *frequent filers* attempted each day to reduce the size of their inbox by filing mails in folders on a regular basis; and *spring cleaners* dealt with overloaded inboxes intermittently or on an irregular basis.[13]

Whittaker and Sidner found in their sample, that 33% of email users were categorisable as *no filers,* 39% *spring cleaners* and 28% *frequent filers.*[14] They also found that *no filers* received more emails per day than *spring cleaners* or *frequent filers* and that filing behaviour appeared to be related to volume of mail received.[15] Using inference testing, they were able to show that this result was statistically significant at the a=0.05 confidence level.

Whittaker and Sidner concluded that email needed to be redesigned to deal with observed problems in filing and task management. Recommended reforms included the use of a common thread ID[16] to enable conversations to be viewed by threads and semantic analysis to cluster semantically related documents automatically as a means of reducing mailbox clutter and the problem of failed folders.

Ten years later, in 2006, two further studies, one emanating from Microsoft[17] and another originating from the Netherlands,[18] revisited Whittaker and Sidner's concern with information overload. Fisher et al, analysed user behaviour as represented by email archives using Microsoft's Social Network Analysis Relationship Finder (SNARF). A sample of 600 users was constructed for this study comprising Microsoft employees. Findings showed:

- An increase in the total number of user folders compared with 1996 (2.8 x 1996); and
- A tenfold increase in the volume of email archives held by users compared with 1996.

Depending on thresholds adopted, relative distributions of *no filers, spring cleaners* and *frequent filers* were either roughly the same as the earlier Whittaker and Sidner study, or showing an increase in *spring cleaners* at the expense of *no filers*. Importantly, percentage distributions for *frequent filers* were consistent between 1996 and 2006, showing that this pattern of behaviour is typical of only between 28% (1996) and 21% (2006) of email users.[19]

In another 2006 study, Janssen and de Poot used critical incident analysis to understand a variety of information overload situations resulting in work-related stress, decreased job satisfaction and poor performance. The study group for this study comprised senior managers working in an industrial company. Scenarios investigated included email cascades and avalanches, email workload, ambiguous email, unwelcome notifications and bad email practices.[20] Subjects were categorised into respondent groups consisting of non-sufferers, permanent sufferers and occasional suffers. Janssen and de Poot found that the extent to which users display symptoms of information overload was related to coping strategies. Coping strategies that emphasise information organisation, such as filing and archiving, were found to be associated with *permanent sufferers.*

Behaviour and volition: reconstructing the locus of intervention

Results from each of these three studies of end-user responses to the problem of information overload sit uncomfortably with the expectations that users will conform with recordkeeping requirements involving the filing of email to corporate stores. If only between 21% (Fisher et al, 2006) and 28% (Whittaker and Sidner, 1996) of email users have a *frequent filer* profile, and up to one third are categorisable as *no filers,* efficient email archiving is unlikely, if driven by users. A well-constructed API interface that reduces, but does not eliminate user burden in filing to the corporate store, is unlikely to change the equation from a user perspective. Observation of end-user filing behaviour therefore sits uncomfortably with archival aspirations. Further, the Janssen and de Poot 2006 study suggests that if archivists and other recordkeeping professionals are prescriptive about user filing behaviour, then this is likely to contribute to information overload resulting in stress and lost productivity.

In short, evidence on filing behaviour shows not only the limitations of policy and procedure type approaches to compliance and email archiving, but also its naïveté. If users are not systematically filing emails, and filing is adding to stress and lost productivity, why do recordkeeping professionals continue to insist that emails be filed to corporate stores, folders or anything else? Turning users into filing clerks is at odds with what is known about user responses to the problem of overloaded inboxes. Further, if overload is increasing and the association between filing behaviour and mail volume suggested by Whittaker and Sidner is true, then the existing tension between user behaviour and assumptions made by recordkeeping professionals can only grow. Ominously, concern about email induced information overload is becoming more common, with the *New York Time*s declaring email a '$650 Billion drag on the economy' and that email is the bane of professional lives.[21]

Evidence based research appears to punch a rather large hole in the good practice credentials of many of our current professional and systems responses to the problems of email archiving and compliance. Understanding how we have come to get it wrong, involves distinguishing *behaviour* from *volition* and their respective consequences in email management. In a typical EDRMS implementation, the system manages artefacts of *volition* i.e. the information products of filing decisions. That works for *frequent filers*, but an effective solution also needs to provide for *no filers* and *spring cleaners* who delay or avoid filing decisions. In other words, an effective solution must be capable of encompassing the full range of behavioural profiles identified with users. So do we have the locus of intervention correct? If not what alternatives exist? Certainly, these questions are being asked in the literature with some authors and practising professionals, such as Iron Mountain Executive Vice President, Brian Murphy, expressing reservations about the wisdom of applying EDRMS solutions to the problem of email archiving.[22]

Shifting the management paradigm from volition to behaviour

If a behavioural approach is adapted to the problem of email archiving, the locus shifts from managing artefacts of volition (created as a consequence of filing decisions) to managing the information artefacts of user messaging behaviour. What benefits does this bring? Independent of the recordkeeping community, information systems (IS) professionals and systems vendors have been working to bridge the gap between user behaviour and compliance goals, including email archiving. This has given rise to a notion in the IS community of vendors and practitioners of 'active email archiving'. According to C Dicenzo and D Smith active email archiving is

> a continuous process that captures and stores, in read-only form, all email sent or received, indexes the email based on header information and, in many cases, does a full text index of the message and attachment content. The captured messages are continuously, or in periodic batches, sent to disk, optical or tape storage. Single-instance store techniques ensure that only one copy of the message is stored and the data is often compressed for storage efficiency. Because some active-archive applications also remove the archived messages periodically from the email system data store and allow users to access those messages in the archive, the active archive is more often disk or optical than tape.[23]

As the description suggests, active email archiving focuses on capturing evidence of behaviour and does not rely upon user intent for message capture. Active archiving can be implemented by capturing messages at the SMTP[24] Internet mail relay or gateway, or by leveraging capture capabilities within the native email application, such as journaling. Figure 2 provides an information architectural view of such a system, inclusive of EDRMS.

Because active email archiving does not depend on end-user filing and sentencing, it is more efficient at capturing record mail with archival significance (as well as non-record mail). When linked to efficient discovery tools, vendors can guarantee the capture of all important email

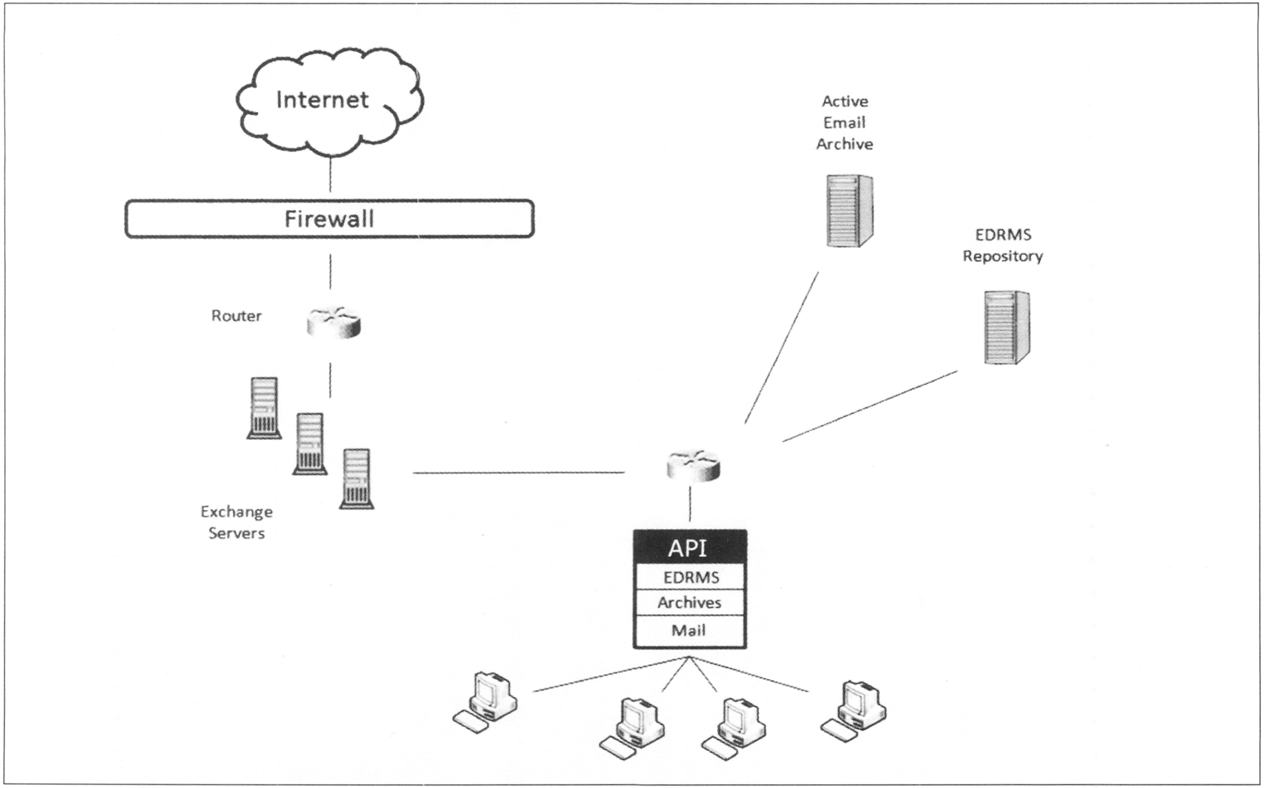

Figure 2. Architectural View of Active Email Archiving

messages into a searchable archive. However, in its crudest implementation, this is not archiving as archivists understand it, but more analogous to backup or replication. In understanding active email archiving and appreciating its strengths and weaknesses, it is important to understand that 'to archive' has a very different meaning for IS professionals. Active email archiving is considered archival by IS professionals because the application involves:

- removal of email from the system data store; and
- writing removed email to high capacity disk, optical or tape storage, using hierarchical storage management (HSM) technologies.

In the IS community, active email archiving is being driven by the growing size of email data stores, regulatory compliance and risk perception. As Swartz observes, judicial activism on 'legal preservation' involving increasing interest in retention and disposal arrangements for documents, as well as records, is a major business driver of email archiving.[25] Gartner estimates that the market, which recorded growth of US$207 million in new software license revenue 2005–2006, will grow to US$1.0 billion by 2011.[26]

Trends in the active email archiving marketplace include new capabilities that enable message archiving to operate on the basis of message value (more in line with the way archivists and records managers see archives management), and improved applications program interface (API) integration providing a more seamless experience for email system users and EDRMS users. Dicenzo and Chin comment that some of these directions are suggesting features seen in records management solutions.[27] An important area of investigation is the development of intelligent systems methods and technologies for machine based sentencing of messages, based on business rules, as originally envisaged by Bearman in 1994. However, while automated classification and indexing has been part of the EDRMS scene for some time, the extension of these ideas into records retention and archiving suggests issues of granularity and risk that may undermine the business logic of email archiving. Correct application of disposal classes is as much a matter of context as content, posing problems for machine based reasoning.

Archival theory and email archiving

IS notions of email archiving jar with mainstream archival thinking about the nature of archives and archival methods. For example, if archiving is regarded as an activity based on records, and records result from documents being set aside by a physical or juridical person, then IS frames of email archiving are at best heretical.[28] However, archivists are by no means uniform in such a view of the record. Noting differences between the Pittsburgh and InterPARES conceptual models, David Bearman recently observed the social risk to the record posed by users who may choose not to 'set aside' as envisaged by InterPARES.[29]

Perceptions of social and other forms of risk have given rise to a view that email governance is important for individual and organisational accountability. A companion notion is that of compliant email management. The idea that organisations must comply with laws, standards and policies developed around the idea of email as a record. In email management, compliance and accountability related risk are therefore most appropriately managed by 'over retention' as Bearman puts it.[30] This over retention leads to a design focus in email archiving on capturing comprehensively the information artefacts of behaviour, that is, documents, rather than records.

Current vectors in systems design for email archiving clearly offer recordkeeping professionals an opportunity to engage IS professionals in a cross-frame dialogue that might result in email archiving solutions that include genuine archiving functionality. For example, email archives are well suited to XML normalisation as a long-term retention strategy. In this sense, the archival systems boundary in email archiving systems has yet to be defined, and can be related to digital preservation discourse on the role of XML normalisation.

Conclusion

Evidence-based research on user email management suggests the limitations of approaches to archival email management that rely on user filing of email.[31] The volume of email received by users leads to the adoption of coping strategies based on delay and the avoidance of decision-making, including filing decisions involving the corporate store. Other evidence suggests that by requiring users to file record mail, recordkeeping professionals may unwittingly contribute to work-related stress and lost productivity.[32] User burden and reduced productivity both work against user acceptance of solutions to the problem of email management that rely upon user filing behaviour.

More effective archival management of email might be achieved by conceptualising the problem of email archiving as one of capturing the artefacts of behaviour, rather than volition. Essentially, this leads to the document as a locus of intervention, something found in the current generation of compliance oriented email archiving solutions. Whether archivists elect to be involved in a cross-frame dialogue with IS professionals or not, greater attention is required in digital recordkeeping to the findings of evidence-based research on user behaviour and attitudes to email management. Systems solutions that contribute to information overload increase social risk to the record and are ultimately self-defeating. Clipping Mercury's wings and managing the messaging mess efficiently will likely require modification of traditional approaches based on volition and 'recordness', in favour of approaches that make sense in terms of what is known about user behaviour.

Endnotes

[1] J Leopold, *White House Official Tells Judge Searching for Lost Emails Requires Too Much Work*, 2008, available at <*http://www.opednews.com/articles/genera_jason_le_080322_white_house_official.htm*> accessed 28 March 2008. See also United States District Court for the District of Columbia, *Citizens for Responsibility and Ethics in Washington (Plaintiff) v. Executive Office of the President, et al.*, 2009, available at <*http://www.citizensforethics.org/files/20090115 – Facciola Memo.pdf*> accessed 15 May 2009.

[2] The Raleigh Chronicle, *Governor Asks For Email Policy to be Reviewed,* (2008), available at <*http://raleigh2.com/default.asp?sourceid=&smenu=1&twindow=&mad=&sdetail=688&wpage=1&skeyword=&sidate=&ccat=&ccatm=&restate=&restatus=&reoption=&retype=&repmin=&repmax=&rebed=&rebath=&subname=&pform=&sc=2502&hn=raleigh2&he=.com*> accessed 28 March 2008.

[3] Corruption and Crime Commission, *Report on the Investigation of Alleged Misconduct concerning Dr Neale Fong, Director General of the Department of Health,* Perth, WA, Corruption and Crime Commission, 2008, available at <*http://www.ccc.wa.gov.au/pdfs/report-alleged-misconduct-fong-neale.pdf*> accessed 22 March 2008.

[4] ibid., p. 30.

[5] See D Wallace, 'Implausible Deniability: The Politics of Documents in the Iran-Contra Affair', in R Cox and D Wallace (eds), *Archives and the Public*

Good: Accountability and Records, Westport, Conn., Quorum Books, 2002, pp. 91–114.

[6] D Bearman, 'Managing Electronic Mail', *Archives and Manuscripts,* vol. 22, no. 1, May 1994, p. 29.

[7] T Allman (ed.), 'The Sedona Conference Commentary on Email Management: Guidelines for Selection of Retention Policy', *The Sedona Conference Journal,* vol. 8, Fall 2007, p. 239.

[8] D Dahlquist, 'Confidence in Email Archiving Waning', CMS Wire website, 14 March 2008, available at <*http://www.cmswire.com/cms/enterprise-cms/confidence-in-email-archiving-waning-002427.php*> accessed 3 June 2008.

[9] Microsoft, *Enterprise Content Management: Breaking the Barriers to Broad User Adoption,* Microsoft White Paper, July 2006, available at <*http://download.microsoft.com/download/3/a/c/3ac5382c-d2dc-4361-881b-8dc3643b8552/ECM_WhitePaper.pdf*> accessed 10 June 2008.

[10] V Bellotti, N Ducheneaut, M Howard, I Smith and R Grinter, 'Quality versus quantity: Email-centric task management and its relations with overload', *Human–Computer Interaction,* vol. 20, nos 1 and 2, June 2005, pp. 89–138; L Dabbish, R Kraut, S Fussell and S Kiesler, 'Understanding email use: Predicting action on a message', in *Proceedings of Computer Human Interaction (CHI),* 2005, pp. 691–700; S Whittaker and C Sidner, 'Email Overload: Exploring Personal Information Management of Email', Proceedings of the Association of Computing Machinery Conference on Computer–Human Interaction (CHI), 1996, pp. 276–283; S Whittaker, V Bellotti and J Gwizdka, 'Email in personal information management', *Communications of the ACM,* vol. 49, no. 1, January 2006, pp. 68–73.

[11] Whittaker and Sidner, p. 280.

[12] ibid.

[13] ibid., p. 280.

[14] ibid., p. 281.

[15] ibid., p. 282.

[16] A unique identifier ID shared by email threads enabling beginning to end reconstruction of a thread.

[17] D Fisher, A Brush, E Gleave and M Smith. 'Revisiting Whittaker and Sidner's "Email Overload" Ten Years Later', *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work,* Banff, Alberta, Canada, 4–8 November 2006, pp. 309–312, available at <*http://research.microsoft.com/pubs/69394/p309-fisher.pdf*> accessed 13 May 2009.

[18] R Janssen and H de Poot, 'Information overload: Why some people seem to suffer more than others', *Proceedings of the 4th Nordic Conference on Human-*

Computer Interaction: Changing Roles, Oslo, Norway, 14–18 October 2006, pp. 397–400.

[19] D Fisher et al, p. 312.

[20] Janssen and de Poot, p. 398.

[21] S Lohr, 'Is Information Overload a $650 Billion Drag on the Economy?', *New York Times,* 20 December 2007, available at <*http://bits.blogs.nytimes.com/2007/12/20/is-information-overload-a-650-billion-drag-on-the-economy/*> accessed 2 September 2008.

[22] B Murphy, 'Preventing Mistakes in Email Records Management', *The CPA Journal,* vol. 75, no. 7, 1 July 2005, p. 14.

[23] C Dicenzo and D Smith, 'What is Email Active Archiving?', Gartner research no. TU-21-5490, 24 November 2003, p. 2.

[24] Simple Mail Transfer Protocol (SMTP).

[25] N Swartz, 'Enterprise-Wide Records Training: Key to Compliance, Success', *Information Management Journal,* vol. 40, no. 5, 1 September 2006, p. 36.

[26] C Dicenzo and C Chin, 'Magic Quadrant for Email Active Archiving', Gartner research no. G00148154, 16 May 2007, p. 4.

[27] ibid., p. 2.

[28] InterPARES, Authenticity Task Force, *Authenticity Task Force Report,* (no date), p. 1, available at <*http://www.interpares.org/book/interpares_book_d_part1.pdf*> accessed 2 September 2008.

[29] D Bearman, 'Moments of Risk: Identifying Threats to Electronic Records', *Archivaria,* no. 62, Fall 2006, p. 44.

[30] ibid., p. 27.

[31] See Whittaker and Sidner; and Fisher et al.

[32] See Janssen and de Poot.