Where Light in Darkness Lies

Preservation, Access and Sensemaking Strategies for the Modern Digital Archive[1]

Jason R. Baron[1] and Simon J. Attfield[2]

[1]US National Archives and Records Administration

[2]Interaction Design Centre, Middlesex University

Abstract

The second decade of the 21st century finds institutions around the world increasingly having to cope with the matter of how to preserve electronic records in response to litigation, regulatory and compliance demands. In the public sector, existing approaches to email preservation in particular (based on past litigation) run the gamut from continued reliance on print to paper strategies; to deployment of disaster recovery backup tapes as default recordkeeping systems; to forms of electronic recordkeeping that continue to rely on end users performing records management functions; and most recently, to automated email archiving. This paper argues that with greater encouragement and adoption of automated capture methods, archivists and historians should understand both the limitations of present-day search techniques, as well as the need to extract "meaning" out of what are increasingly vast amounts of electronic records, especially through consideration of new methods drawn from the disciplines of sensemaking and visual analytics.

Authors

Jason R. Baron serves as Director of Litigation for the US National Archives and Records Administration, and is an Adjunct Faculty Member at the University of Maryland, College of Information Studies. Mr. Baron formerly held the positions of trial attorney and senior counsel in the U.S. Department of Justice. He has also served as a Visiting Scholar at the University of British Columbia, participated in InterPARES, and co-founded the US NIST TREC Legal Track. He is a recipient of the Emmett Leahy Award, recognizing his career contributions in the field of records and information management. Mr. Baron received his degrees from Wesleyan University and the Boston University School of Law.

Simon Attfield is a Senior Lecturer in the Interaction Design Centre, Middlesex University, UK. His research lies in the area of understanding how people work with information, processes involved in individual and collaborative sensemaking and implications for interactive systems design. With Ann Blandford he is co-author of the book *Interacting with Information*, part of the Morgan Claypool series of Synthesis Lectures on Human-Centered Informatics. He received a B.A. in Philosophy and a BSc. in Experimental Psychology from Sussex University, and a PhD in Human Computer Interaction from University College London.

---

[1] This article represents a reworking of themes and materials presented at the 8th European Digital Preservation Conference (ECA 2010) in Geneva, at I-CHORA 5 (2010) in London, as well as the DELOS conference (2007) in Rome. The authors wish to thank Richard Cox, Richard Pierce-Moses, Mary Jo Pugh, and Gary M. Stern for their comments on prior drafts, as well as Luciana Duranti for her overall support. The views expressed here are solely the authors, however, and do not purport to represent the official position of any institution with which they are affiliated.

1. Introduction

During the past 20 years, with the growth of computer networks and interconnectivity in the workplace,[2] email has become a worldwide phenomenon, transforming the lives of individual employees. The replacement of secretaries with personal computers has, for better or worse, turned each office worker into a *de facto* record keeper in public bureaucracies, and to a lesser extent, in private firms and corporations as well. Now, at the start of the second decade of the 21st century, public and private institutions around the world are increasingly coping with the consequence of having empowered all staff with email as a communications tool, viewing email as a corporate necessity. Paradoxically, *de facto* - modern email archives remain a major source of risk, given the minefield represented by modern day litigation, investigations, compliance and regulatory measures, all in search of evidence in the form of a candid "smoking gun." And yet, the email universe continues to expand exponentially, giving rise not only to short term information governance challenges, but also profound issues regarding access to the contents of those communications -- including the desire to find relevant messages as well as in making sense of the collection as a whole.

Public sector institutions have deployed variety of policy and technology solutions as "preservation" policies: continued reliance on print to paper strategies; deployment of disaster recovery backup tapes as default recordkeeping systems; electronic recordkeeping that relies on end-users to "tag, drag and drop" individual email communications into an archive; and most recently, tentative adoption of automated capture methods known as "email archiving." All these technological solutions, other than "print to paper," more or less rely on automated searching to provide access to email archives in their electronic form. By now, however, well-established limitations on the efficacy of keyword searching leave open for resolution how best alternative strategies may be employed for extracting "meaning" from vast amounts of email, i.e., how to perform information retrieval to extract the "needles" of meaning from the ever growing e-haystack.[3]

There is, however, a fundamental "records management" reality that needs to be confronted and overcome. Even two decades into the deployment of email, and even in the highly networked world we find ourselves in, the reality of the "end user's" experience in managing and preserving electronic records has not fundamentally changed. In particular, in the United States at present, only a few public sector institutions have instituted successful, comprehensive, enterprise-wide electronic content management, even in the face of sustained criticism that the state of records management in the federal sector in the U.S. approaches "chaos."[4] Change is now in the air, however, as most notably evidenced in new records management mandates from President Obama and U.S. Archivist David Ferriero.

---

[2] Martin Hilbert and Priscilla Lopez, "The World's Technological Capacity to Store, Communicate, and Compute Information," *Sciencexpress (*Feb. 10, 2011) (in 2007 humankind was able to store 290 exabytes and the rate is increasing by 23% a year). Accessed October 9, 2012, http://www.sciencemag.org/content/332/6025/60.

[3] See The Sedona Conference® Best Practices Commentary on the Use of Search and Information Retrieval Methods Used in E-Discovery," 8 *Sedona Conf. J.* 189 (2007), www.thesedonaconference.org/publications.

[4] ., Citizens for Responsibility and Ethics in Washington (CREW), "Record Chaos: The Deplorable State of Electronic Record Keeping in the Federal Government," (2008), http://www.scribd.com/doc/49055979/Record-Chaos-Report; U.S. Government Accountability Office, "Federal Records: National Archives and Selected Agencies Need To Strengthen E-Mail Management," GAO-08-742 (2008), http://www.gao.gov/new.items/d08742.pdf.

2. A New Day for Managing Public Sector Email & A Challenge

On November 28, 2011, President Obama issued a memorandum on Managing Government Records for the heads of all Executive branch agencies in the U.S., stating that he was beginning a government-wide "effort to reform records management practices."[5] He recognized that:

> Decades of technological advances have transformed agency operations, creating challenges and opportunities for agency records management. Greater reliance on electronic communication and systems has radically increased the volume and diversity of information that agencies must manage. With proper planning, technology can make these records less burdensome to manage and easier to use and share. But if records management policies and practices are not updated for a digital age, the surge in information could overwhelm agency systems, leading to higher costs and lost records.

In stating that "proper records management is the backbone of open government,"[6] the President directed that the Archivist of the United States, working with other senior officials, develop a Records Management Directive ("Directive") that addresses modern day records challenges, including with respect to email, social media, and cloud based data sets.

On August 24, 2012, the Archivist of the United States issued the planned for Directive, a document that represents an inflection point in the history of archival institutions worldwide. The Directive boldly sets out that by 2019 all permanent electronic records of the U.S. Government are to be managed and preserved in electronic form, for eventual transfer and accessioning at the National Archives and Records Administration ("NARA").[7] As a corollary, the Archivist also directed that *all* email records (both permanent and temporary in nature), be managed in an accessible electronic format by the end of 2016. Specifically with respect to email, the Directive goes on to say:

Email records must be retained in an appropriate electronic system that supports records management and litigation requirements (which may include preservation-in-place models), including the capability to identify, retrieve, and retain records for as long as they are needed.[8]

The Directive separately calls for NARA and others to work with private industry "to produce economically viable automated records management solutions," including "advanced search techniques,"[9] and to "embed records management requirements into cloud architectures."[10] These promising mandates at long last give a much needed impetus to finding satisfactory electronic archiving solutions for email and other forms of electronic records. They represent a profound commitment to true digital government: that what is born digital should be preserved digitally, at least insofar as the data, information and records represent what is deemed appropriate to permanently preserve the memory of the 21st century.

But these aspects of the Archivist's directives represent a challenge to archivists, as well as to all those seeking access to archives (including lawyers, historians, and researchers) – given the new reality that we are living in a world of vast public sector digital archives that are expected to be sustained on an

---

[5] *See* http://www.whitehouse.gov/the-press-office/2011/11/28/presidential-memorandum-managing-government-records.
[6] Ibid.
[7] *Id.,* Part I, § 1.1.
[8] *Id.,* § 1.2.
[9] *Id.,* Part II, § A3.
[10] *Id.,*§ A4.

indefinite basis into the long-term future. Compliance with the new Directive requires the management and ultimately long-term preservation of permanent record emails in digital form. Given this mandate, archivists in public sector institutions such as NARA also have a responsibility to be good stewards of these growing digital archives in providing access. Thus, beyond resolving difficult preservation issues, greater attention also needs be paid to the conduct of search and sensemaking within these large digital collections, for purposes of best ensuring or optimizing future access as expeditiously as possible, within the constraints of the law, for the benefit of future generations.

In the United States, due to litigation involving White House email that began on the last day of the Administration in Ronald Reagan (in January 1989), presidential records in the form of email are deemed permanent public records and have been preserved.[11] By 2017 there may be as many as one billion White House emails cumulatively in the legal custody of the Archives.[12] The White House email archiving experience against the backdrop of continuing litigation lasting over two decades amounts to a tale both fascinating and cautionary, with a mixed bag of lessons.[13] That this archive has been continuously maintained over two decades, across a variety of proprietary platforms -- with whatever their admitted defects -- serves to indicate that automated solutions to email archiving are indeed feasible.

Yet, paradoxically, with limited exceptions all email that has come into the National Archives to date remains behind closed doors, essentially locked away in the digital dark. Litigation and Freedom of Information Act requests provide limited forays into these dark places, but they do not represent the kind of systematic review that is needed to open these archival collections in a logical way. NARA's presidential libraries are overwhelmed with lengthening access queues due to both litigation and filings under the Freedom of Information Act. However, because the collections do contain sensitive and privileged information -- everything from social security numbers to medical conditions to prior criminal records and the like -- they cannot easily be made open to the public without further archival processing. This in turn means that large segments of these email repositories exist as *de facto* digital dark archives, awaiting future systematic opening in whole or in part determined by the longest period covered by existing federal privacy laws and practices. Under such circumstances, the default condition on these archives is that they will not become open to the public until 75 or more years after the end of the current Administration.[14]

In light of all of the above, a commitment to sustainable digital archives necessitates a parallel commitment to new ways of thinking: first, about the burden on end-users in performing records

[11] See Armstrong v. Executive Office of the President, 1 F.3d 1274 (D.C. Cir. 1993).

[12] See George L. Paul and Jason R. Baron, "Information Inflation: Can The Legal System Adapt?," Richmond J. Law & Tech. 13, no. 10, ¶ [15] (2007), http://law.richmond.edu/jolt/v13i2/article10.pdf" (estimating 1 billion White House emails in the legal custody of NARA by 2017).

[13] See Jason R. Baron, "The PROFS Decade: NARA, Email and the Courts," chap. 6 in Bruce I. Ambacher, ed., Thirty Years of Electronic Records (Lanham, MD and Oxford: The Scarecrow Press 2003); David Bearman, "The Implications of Armstrong v. Executive Office of the President for the Archival Management of Electronic Records," American Archivist 56, no. 4 (1993): 674-89; GAO Report 01-446, "Clinton Administration's Management of Executive Office of the President's E-mail System (April 2001), http://www.gao.gob/products/GAO-01-446; Citizens for Responsibility and Ethics in Washington, "The Untold Story of the Bush White House E-mails" (August 2010), http://citizensforethics.org/bush-white-house-ignored-warnings-about-email; Myron Groover, "The White House E-Mail Destruction Scandal of 2007: A Case Study for Digital Heritage," presented at the UNESCO Conference: The Memory of the World in the Digital Age, Vancouver, B.C. (2012).

[14] See NARA regulations at 36 C.F.R. 1256.56(a)(2) (2012) (NARA policy potentially restricts information in personnel, medical or other records revealing information of a highly personal nature regarding living individuals that inter alia "relates to events less than 75 years old.").

management; second, about the need for automating the segregation or classification of permanent records apart from the disposable or transitory, and third, about the opportunity to use advanced search techniques to extract meaning and make sense of the coming vast digital collections.

3. Past Failed Approaches to Email Management

We first need to declare an end to the era of end-users assuming the burden of records management. Since the late 1980s, with the universal rollout of PCs in offices and at personal workstations, all end-users have become *de facto* records managers, which may have been theoretically achievable in a world of word processing, spreadsheets, and the *occasional* email of substantive importance. However, the reality of the present day workplace is vastly different: each end user is drowning in information, only a portion of which in the public sector constitute long-term temporary or permanent records.

Except in certain exotic and rarefied environments, print-to-paper appears to be an utterly failed recordkeeping paradigm. Substantial rates of individual end-user noncompliance with retention policies calling for the printing out of email are the norm in any organization that will admit to such in a survey (or under oath). The U.S. Government Accountability Office reported in 2008 on a small survey of 15 senior officials at various federal agencies, eight of whom "did not consistently conform to key requirements in NARA's regulations for email records, such as filing them in appropriate recordkeeping systems."[15] The simple truth: with rare exception, for the past twenty years, few professionals (and even fewer lawyers) consistently or comprehensively have printed out all emails of record for placement in traditional office filing systems.

But other more digital-centric recordkeeping approaches that nevertheless rely on end-users have similarly met with less than success. Strategies that rely on ad hoc actions on the part of individual end-users to archive email into personal folders do not constitute comprehensive recordkeeping of value to the organization as a whole, and these methods flunk the current test under federal e-recordkeeping guidelines for what constitutes an electronic recordkeeping system.[16] At the same time, some percentage of public sector organizations continue to store email on disaster recovery backup tapes with minimal or no recycling involved, as their electronic recordkeeping "solution." Agency officials are often oblivious to the distinction between "archiving" and "recordkeeping" functions, as well as to the difference between each of those and saving data to disaster recovery backup tapes – thus policies relying on backup tapes may not be seen by them as inherently problematic. Email and other e-records "stored" on these forms of backups, however, have been physically streamed in no logical order and must be "restored" by a labour-intensive process of remounting sets of backups from a single day or session, to extract desired files from reconstructed user accounts. Disaster recovery backup tapes are not databases or logical repositories of information and cannot be easily searched without great time, money and effort.[17] That is why NARA

[15] *See* GAO Report 08-742: 4. *See also* NARA, Records Management Self-Assessment (2009): An Assessment of Records Management Programs in the Federal Government, http://www.archives.gov/records-mgmt/resources/self-assessment.html; SRA International, Inc., Report on current Recordkeeping Practices with the Federal Government (Dec. 10, 2001), http://www.archives.gov/records-mgmt/pdf/report-on-recordkeeping-practices.pdf.

[16] *See* 36 C.F.R. 1236.20 (requiring records disposition functionality and shared access).

[17] *See generally* Grant J. Esposito and Thomas M. Mueller, "Backup Tapes, You Can't Live With Them and You Can't Toss Them: Strategies for Dealing with the Litigation Burdens Associated with Backup Tapes Under the Amended Federal Rules of Civil Procedure," Richmond J. L. & Tech. 13, no.13 (2006), http://law.richmond.edu/jolt/v13i3/article13.pdf.

states in its federal recordkeeping regulations that backups "must not be used as the agency electronic recordkeeping system."[18]

A cottage industry of records management applications (RMAs) by way of software products and services also exists. These products and services provide organizations with the capability of empowering end-users with greater control of the ability to perform electronic recordkeeping. The most visible standard for e-recordkeeping is the 5015.2 Standard (now in version 3 form) employed by the U.S. Department of Defense since 1997, which sets out several dozen functional requirements for RMA software to meet in order to demonstrate compliance with federal recordkeeping standards.[19] A significant number of software products have been certified as 5015.2 compliant, and NARA has encouraged their use.[20]

For all of the support given during the past decade of this "first wave" of 5015.2 compliant RMA applications, they have not been widely deployed to date within government agencies. Part of the problem may be residual difficulty in scaling up the software that comes out of the box, to meet the needs of particular large-scale institutions and enterprises. However, what we perceive to be the much larger issue is the transactional demands that this form of software makes on end-users, the vast majority of whom do not consider records management a top priority as they go about their daily workplace tasks. RMAs modelled on 5015.2 version 3 normally require users to select how individual emails and other e-records are to be managed from drop-down screens. Often times, institutions have not thought through how they would simplify their existing stock of legacy records schedules, so as not to impose a bewildering array of choices on users indicating the series in which individual electronic objects are to be placed. The result in some cases has been failed pilot projects, where users simply rebelled, or in the more usual case, underutilized the RMA application due to the transactional burden involved. The jury measuring the success of RMA applications is still out, however: it may well be that the product lines need further maturation, that the end-user experience will be easier, and that this form of e-recordkeeping solution might eventually become more pervasive than it is currently.

Anecdotal experience with DoD Standard 5015 RMA applications has shown that many individuals do not reliably save, tag, drag or drop most of their email into electronic folders set up with well-intentioned archival purposes in mind. The compliance rate starts off low in this regard, and is invariably getting lower – simply due to massive volumes of information now confronting us all. Having to perform extra keystrokes, no matter how few, on any substantial percentage of electronic communications during the work day, means the equivalent of having to pay a transactional toll per communication, in terms of lost time, energy, and productivity. Few individuals wish to pay the price on a consistent and comprehensive basis, hence, a world of incomplete if not haphazard recordkeeping.

Aside from the inherent problems of backup tapes, the Achilles heel for the remaining above-described methods should be plain: that is, the extent to which these methods rely on individuals to perform recordkeeping ("self-archiving") at the desktop. Given the reality of the sheer volume of email

---

[18] 36 C.F.R. 1236.20(c) (2012).

[19] *See* http://www.archives.gov/records-mgmt/policy/joint-interoperability-letter.html; *see also* http://jitc.fhu.disa.mil/recmgt/.

[20] NARA Bulletins 2003-03 and 2008-07, "Endorsement of DoD Electronic Records Management Application (RMA) Criteria Standard, versions 2 and 3, respectively, *see* http://www.archives.gov/records-mgmt/bulletins/2003/2003-03.html and http://www.archives.gov/records-mgmt/bulletins/2008/2008-07.html

and other electronic records coming across the transom every day, coupled with the transactional cost of compliance, sole dependence on software that requires diligence by end-users is tantamount to institutions whistling past the graveyard, at least with respect to their immediate e-discovery obligations, and with downstream recordkeeping implications absent strong training and enforcement programs in place. In our view, agencies have a responsibility to look for alternative strategies in light of these real-world considerations.

4. The Promise of Email Archiving Through Automated Capture

As distinct from the above approaches, it is now increasingly apparent that software technology has advanced significantly to the point that we can speak in terms of something truly new: automated email archiving for the enterprise. Such a notion of comprehensive capture of email by automated means conjures up Borgesian visions[21] of near-infinite library repositories – a dream to some, a nightmare to others, including perhaps appraisal archivists grounded in longstanding notions of what constitutes the "judicious" disposition of records.[22] At least today, email archiving software starts with the promise of the comprehensive capture of email, not just as selected by users, but as pulled from the underlying proprietary email system on a continual basis. The concept of "automated email archiving" is predicated on the notion that 100% of email traffic is captured or routed into some form of online or near-line electronic environment. Without the type of fanfare accompanying iPhones and other 21st Century new age gadgetry, this new form of software and services nevertheless has gained a wide audience, and is currently being piloted in a variety of both public and private sector institutions as a way to manage risk, including the risk associated with having to comply with litigation demands for "all" e-communications.

Automated email archiving is wonderfully simple in its conception: email, and whatever else, is captured from whatever email proprietary system or email store is in use on the online server, with the electronic object then placed in a separate database either near-line or offline to be preserved in its native proprietary form under whatever rule set is proscribed. The electronic objects are therefore accessible from a central location. Indeed, the system may well be set up that the user is able to "see" or otherwise freely access such archived email in the separate database as if the email still resided in his or her original in-box (through the use of what is referred to as an email "stub"). From all of the above discussion, it should be clear that the value of such systems for e-discovery purposes is apparent, in lowering the risk of lost records while increasing accessibility through means of expedited searches.[23]

As NARA's bulletin on the subject of email archiving pithily states, "Email archiving applications typically require little to no action on the part of the user to store or manage the email records."[24] NARA recognized that benefits include (i) more efficient storage of email because it is moved from a distributed network of servers and desktop applications into one place; (ii) enhanced search capabilities for content germane to subpoenas, FOIA requests, and e-discovery requests; and (ii) assistance in backup and disaster recovery.[25] NARA's bulletin goes on to note, however, that "[e]-mail archiving is a relatively new

---

[21] Jorge Luis Borges, "The Library of Babel," Labyrinths: Selected Stories and Other Writings (1971): 51.
[22] Title 44 of the U.S. Code, Section 2902(5) (listing "judicious preservation" as one of the objectives of records management).
[23] *See* Wikipedia, "Email archiving." Accessed October 9, 2012, http://en.wikipedia.org/wiki/Email_archiving.
[24] NARA Bulletin 2011-03, "Guidance concerning the use of e-mail archiving applications to store e-mail" (December 2010), http://www.archives.gov/records-mgmt/bulletins/2011/2011-03.html.
[25] NARA Bulletin 2011-03: 2.

technology and is still developing," and that agencies must appropriately configure and implement records management controls when using the application.[26]

We believe, however, that solutions already exist in the marketplace that meet some level of the recordkeeping controls that NARA would wish to see in place. For example, consistent with existing records retentions schedules, an agency may decide to implement an automatic rule "tagging" all email created or received by designated senior officials as "permanent." Remaining email from all other agency staff could be categorized as presumptively "temporary" in one or more "big buckets" or folders within an email archiving scheme corresponding to existing record series. Alternatively, a similar tagging rule could be employed for all email created or received by designated components that routinely generate high-level policy or other sensitive documents worthy of long-term preservation.

An agency might elect any number of "add-on" measures, to bring greater nuance to the archiving scheme. These might include simple steps aimed at allowing for users to create individual folders within the email archive, for the limited purpose of ensuring that some measure of "virtual foldering" takes place akin to traditional records management in the paper world. For senior staff with designated permanent records, agencies concerned about over-inclusion of email of a personal or truly ephemeral nature could allow for staff to delete emails from the in-box for a limited period of time (e.g., 60 or 120 days), with any emails remaining then automatically captured as permanent under the "auto-archiving" rule in effect. Conversely, mid- or lower level staff who do create "permanent" records could be allowed a manual override of the "temporary" designation, so as to "opt-in" for records preservation on individual or designated categories of email sent or received.

In marketing email archiving solutions, much has been made of the notion that the software also generally allows for greater records management functionality. One of the multiple reasons why organizations implement email archiving is said to be to meet litigation, regulatory, and/or business records retention requirements, by enabling easier searches of stored email. On closer examination, there appear to be no perfect, scalable solutions yet to truly automating a large organization's email by content, even using techniques borrowed from the world of information retrieval, data mining, and artificial intelligence. Some vendors in the marketplace are, however, touting the ability to have systems intelligently "auto-classify" or "auto-categorize" documents based on some measure of machine learning from examples provided by users themselves.[27] Such techniques would serve as proxies to more traditional means of recordkeeping that those of us born in the 20th Century are used to in the workplace. Even now, however, the technologies available in the marketplace go a serious way towards meeting the long-anticipated goal of some archivists in being able to adopt more sophisticated forms of macro-appraisal and "institutional functional analysis," for reliably segregating content by function or activity of an organization,[28] and thus are worthy of serious further examination.

---

[26] Ibid.

[27] See Jason R. Baron, "Law in the Age of Exabytes: Reflections on 'Information Inflation' and The Current State of E-Discovery Search, Richmond J. of Law & Tech. 17, no. 9 (2011) , http://jolt.richmond.edu/v17i3/article9.pdf.

[28] Terry Cook, "Archival Appraisal and Collection: Issues, Challenges, New Approaches" (1999), http://www.mybestdocs.com/cook-t-nara-990421-2.htm; Terry Cook, "Electronic Records, Paper Minds: The revolution in information management and archives in the post-custodial and post-modernist era," Archives and Manuscripts, vol. 22, no. 2 (1994); see also Christopher Tomer and Richard J. Cox, "Electronic Mail: Implications and Challenges for Records Managers and Archivists," The Records & Retrieval Report, 8, no. 9 (1992): 7 (referring to a "macro-appraisal/documentary probe" strategy for email).

5. Limitations of Present-Day Search Methods and A Path Forward

To date, the majority of products and services touting electronic archiving solutions appear to have put their eggs in the back-end search basket, rather than implementing more traditional front-end records management solutions. This suggests that search capabilities are robust enough to replace all other notions of provenance, original order, and the need for any form of granular auto-categorization by record series type, features of a repository that archivists have traditionally expected and demanded. In light of the claims being made, it remains important to accurately measure success in performing robust information retrieval. This is, however, a rapidly evolving field, and alternative solutions to Boolean searching being deployed in present day litigation may yet have important implications, including for members of the archival profession.

To the extent email archiving schemes rely on sorting by means of present-day "search" technologies to perform records management-like classification, they are of course on somewhat shaky ground. As is well known in the information retrieval community,[29] the idea that any form of text retrieval algorithm can reliably parse text such as email, based on content alone, into that which is valuable and worth keeping, while setting aside that which is ephemeral and to be disposed, still is viewed as a hard problem -- especially if the goal is perfect sorting of records into 20th Century-style, highly granular records schedule buckets. There are, however, a number of products and services in the e-discovery market that take advantage of clustering algorithms and "predictive" analytics, in ways that would greatly advance the auto-categorization task.

This subject has been explored at length in connection with e-discovery issues, through publications including The Sedona Conference® *Best Practices Commentary on the Use of Search and Information Retrieval Methods Used in E-Discovery*.[30] The Commentary describes at length the limitations involved in present-day search methods based on keywords and Boolean operators, and suggests that lawyers (and others) should be looking in the short term to alternative forms of search methods. The latter methods now include forms of supervised learning (widely known today in the legal community as "predictive coding"), that makes use of the coding of seed sets of documents, coupled with use of statistical or probabilistic forms of searching based on mathematical clustering models, for the purpose of enabling software to in turn code the vast majority of documents in a given repository.[31] Given the exponential increases in volume of electronically stored information, lawyers are increasingly taking advantage of the well-known gains in efficiency in relying on these software-assisted methods, which in turn utilize "iterative" feedback loops involving greater transparency in divulging statistical samples to those inquiring about relevant documents.[32]

For example, in the prominent *da Silva Moore* case decided in 2012, a U.S. magistrate judge held that the state-of-the-art in advanced search techniques had progressed to the point where the Court could "bless" the use of the form of supervised learning known as "predictive coding."[33] The court determined the use of predictive coding was appropriate considering "(1) the parties' agreement, (2) the vast amount

---

[29] For citations to the relevant literature, *see* Douglas W. Oard and William Webber, "Information Retrieval for E-Discovery," Foundations and Trends in Information Retrieval, 6, no. 1 (forthcoming in 2012):1-144. Accessed October 9, 2012, ,http://ediscovery.umiacs.umd.edu/pub.html.

[30] See n.3, supra.

[31] Baron, "Law in the Age of Exabytes."

[32] Paul and Baron, "Information Inflation," ¶ 50.

[33] *Moore et al. v. Publicus Groupe SA*, 2012 WL 607412 (S.D.N.Y. Feb. 24, 2012) (Peck., M.J.), aff'd, 2012 WL 1446534 (S.D.N.Y. April 26, 2012) (Carter, J.)

of electronically stored information to be reviewed (over three million documents), (3) the superiority of computer-assisted review to the available alternatives (*i.e.,* linear manual review or keyword searches), (4) the need for cost effectiveness and proportionality . . . ; (5) the transparent process proposed by [defendants]."[34] The opinion included at its end an extraordinarily detailed "joint protocol" worked out by the parties, setting out the construction of seed sets, sampling, and the seven rounds of iterative training of the system for purposes of classifying documents as responsive or nonresponsive. The opinion points to a future of judicially-blessed litigation engagements where supervised learning in lieu of keyword searching is routinely utilized.

Additionally, over the past half decade the limitations of competing information retrieval methods in a legal context have been explored as part of what is known as the "TREC Legal Track," an international research project sponsored by the U.S. National Institute of Standards and Technology, which ran between 2006 and 2011. The goal of the legal track was to evaluate how alternative forms of search methods perform in a "real-life" situation involving real document databases (consisting to date of seven million tobacco litigation documents, plus the Enron email collection), and a set of hypothetical topics which could have been the subject of e-discovery demands. In other words, the track coordinators wished to draw comparisons as between how well Boolean methods perform against fully automated alternatives proposed and used by scientist-participants in the track. The results of the second year of the Legal Track were eye-opening (at least to lawyers, as opposed to information retrieval researchers): only 22% of the total number of relevant documents were found by traditional keyword or Boolean search methods; 78% were found by all other search methods combined.[35] Results from subsequent years have validated the above numbers while also exploring the efficacy of greater use of human-in-the-loop solutions, as well as supervised learning methods.[36]

The research and analysis represented by the TREC Legal Track points to the difficulty in making very strong claims regarding the efficaciousness of how reliably searches, conducted against a large electronic archive with many millions or billions of objects embedded within it, can successfully classify or segregate information categories. As stated, these methods are rapidly improving, however. Arguably, parsing documents into categories of "relevance" is more difficult than attempting to match documents into relatively broad recordkeeping categories (by series, for example), although any attempt to categorize documents based on content analytics alone will achieve less than perfection. (Of course, some measure of success can be obtained if an organization is willing to set big-bucket rules to designate as permanently valuable "all" records generated by particular individuals or components within the enterprise, without regard to content, as discussed, *supra*.) The bottom line is that while we should remain cautious about naively buying into overly robust assertions regarding the efficacy of searches, there is rapidly emerging progress in the area of search methods and technologies that archivists and lawyers would be well served to remain aware concerning. The promise of the future is that by employing greater use of sampling and iterative techniques, coupled with new forms of supervised learning, e-record archives can potentially be filtered for other purposes -- including for performing macro appraisal and disposition, and for

---

[34] Ibid.

[35] See "Discovery Overload," Law Technology News (2008):36, http://commonscold.typepad.com/eddupdate/2008/01/edd-showcase-di.html; Stephen Tomlinson, et al., *Overview of the TREC 2007 Legal Track*, http://trec.nist.gov/pubs/trec16/t16_proceedings.html.

[36] Oard and Webber, "Information Retrieval for E-Discovery"; see TREC Legal Track Overview papers, http://trec-legal.umiacs.umd.edu/.

determining where resources are to be allocated to find subsets of electronic repositories that may be made available for access in the nearer term.

There are many approaches to search and information retrieval, only some of which have been explored by those in the legal domain. We turn next to the issue of extracting meaning or "sensemaking" from email archives, an area which holds great promise as it becomes better understood.

6. Sensemaking and the Email Archive

In recent years interest in the study of sensemaking has increased, and one area where this has been evident is in the study of people working with electronic information. If we can understand how people make sense of things and how that can best be supported, we are in a better position to design systems to help sensemaking along, and so help people to engage effectively with increasingly large amounts of electronic content. As both a research area and a practical issue, this interest can be said to be motivated by a need for perspectives that link people's moment-by-moment interactions with information, such as searching, gathering, extracting, and structuring, with the internal processes of theorizing, interpreting and understanding.

A number of theoretical perspectives on what sensemaking is and what happens during sensemaking have emerged.[37] A theme that commonly runs through these accounts is the observation that sensemaking involves a reciprocal interplay between bottom-up exposure to information on the one hand, and the top-down interpretation, structuring or theorizing about information on the other. Each occurs during sensemaking, and each affects the other. Interpretation gives meaning to information and even defines what counts as information (as opposed to 'noise'). But it is the information that gives support to or challenges the interpretation, perhaps forcing it to be modified or abandoned in favour of another. Sensemaking has been described as a process of placing stimuli into some kind of framework (e.g. a mental narrative, a model of others' motives and intentions, a spatial "map," etc.), which then allows us to "comprehend, understand, explain, attribute, extrapolate and predict."[38]

The co-dependence between information and interpretation in sensemaking presents something of a paradox. Appreciating what information is relevant to an endeavour depends on interpretation which itself depends on information. We equate this aspect of sensemaking to the idea of the hermeneutic circle. Hermeneutics concerns the theory (or art) of interpretation.[39] Originally concerned with the interpretation of written texts, contemporary hermeneutics has come to include the analysis of interpretive processes in general. According to the hermeneutic circle, parts of a message can only be understood in terms of an understanding of the whole, and yet an understanding of the whole can only arise from an

---

[37] *See, e.g.,* Karl Weick, Sensemaking in Organisations (London; Sage, 1995); Brenda Dervin, "An Overview of Sense-making Research: Concepts, Methods, and Results to Date" (paper presented at the International Communications Association Annual Meeting, Dallas, May, 1983); Gary Klein, et al., "A Data-frame Theory of Sensemaking" in Expertise Out of Context: Proceedings of the Sixth International Conference on Naturalistic Decision Making, Pensacola Beach, Florida, May 15-17, 2003, ed. Robert Hoffman (Lawrence Erlbaum Associates Inc, 2007):113-155; Peter Pirolli and Stuart Card, "The Sensemaking Process and Leverage Points for Analyst Technology as Identified Through Cognitive Task Analysis," paper presented at the International Conference on Intelligence Analysis, McLean, VA, May 2-3, 2005, https://analysis.mitre.org/proceedings.

[38] William H. Starbuck and Frances J. Milliken,"Executives' Perceptual Filters: What They Notice and How They Make Sense," in The Executive Effect: Concepts and Methods for Studying Top Managers, Donald C. Hambrick, ed. (Greenwich, CT, JAI Press, 1988), 35-65.

[39] "Hermeneutics" and "interpretation" have a shared etymology. *See* Lawrence K. Schmidt, Understanding Hermeneutics (Stocksfield, UK: Acumen, 2006).

understanding of the parts[40]. Take a sentence—individual words, which individually are susceptible to multiple interpretations, have their interpretation fixed through an interpretation of the sentence as a whole. And yet, the attributed meaning of the sentence whole depends upon how each individual word is interpreted. This scales-up—the interpretation of a message, such as an email, depends upon some theory of the context in which the message originated (e.g., time, culture, language, personalities, roles, motivations, recent activities and relationships)—and yet an understanding of these depends upon the interpretation of many other such messages.

We illustrate this with an example from a study into the ways in which lawyers makes sense of large collections of emails during some e-discovery investigations. Attfield and Blandford[41] reported an interview study conducted with corporate lawyers who were engaged in a e-discovery investigations. One group of lawyers was investigating potential fraud in the contracts that a company had with another. They conducted keyword searches over a set of recovered documents (primarily emails) and the results were retrieved so that they could manually review them. However, an interpretation of actions in the emails depended upon an understanding of the broader temporal backdrop against which it occurred. For example, the judgment of whether a communication between two parties could be interpreted as bid rigging would depend on when the communication had taken place in relation to a bid lifecycle. But this context was initially unknown, and needed to be constructed from many other similar emails. People may meet, exchange information or even money, but the meaning of these actions, and, important for this example, their meaning in legal terms is fixed by a broader context. Hence, a bootstrapping process must occur at great time and cost. As one senior lawyer said, "the scope of what you're trying to do is immense and you're having to define it as you go along."

Sensemaking involves moving from the part to the whole and back, gradually building an understanding. For the philosopher Schleiermacher, the impasse of the hermeneutic loop can be broken by an initial cursory reading of an entire text followed by detailed readings of specific parts and relating these to the whole. Schleiermacher, however, did not have large email collections to investigate. A cursory reading of thousands to millions and even tens of millions of documents is not possible.

There is, however, an emerging technology, which offers an opportunity for addressing the hermeneutic loop. Visual analytics is the science of analytical reasoning supported by interactive visualizations. At its heart is the idea of interactive graphical representations of large datasets which are, in principle, easy to interpret and provide context from which to engage in more detailed investigation. Visual analytics has at its core interactive visualizations which utilize the high bandwidth of the human visual system to enable us to draw insights from large, complex datasets. Computerized visualization is not new, of course, but with the concept of Visual analytics it becomes part of a more holistic study of computer-supported sensemaking. The form of the solution is to perform automated analysis (of some type) over a document collection and to use the results to structure visual displays that offer insights into the underlying collection. Visual overviews abstract away from data, prioritizing some aspects of underlying content and hiding others.

Visual analytic tools are a relative newcomer to the e-discovery stage, but given the importance of email to this domain, the interactive visual exploration of email has become a significant area of

[40] Schmidt, Understanding Hermeneutics.
[41] Simon Attfield and Ann Blandford, "Discovery-led Refinement in E-discovery Investigations: Sensemaking, Cognitive Ergonomics and System Design," Artificial Intelligence and Law 18, no. 4 (2010): 387-412.

interest[42]. A number of approaches to the visualization of email collections have been offered. It is not our aim to review them here and the interested reader might refer to a review by Lemieux and Baron.[43] We do, however, refer to one example by way of illustration of the approach.

A research system called Enronic by Heer[44] is shown in figure 1. Enronic is an example of a network graph visualization tool. The system takes an email collection and graphically displays connections based on who sent emails to whom. An advantage of generating social networks from email archives lies in the possibility of calculating and social network metrics and using these as display variables. For example, Enronic calculates *connectivity* (often referred to as *degree centrality*) for each node. This is a measure of the number of connections emanating from a node and can potentially indicate a person's power or influence within the network. Various forms of filtering are typical within Visual analytic systems. Enronic allows the user to filter the display for nodes with higher connectivity. Users can also apply an algorithm to hierarchically cluster the network into community substructures and interactively view them in terms of different levels of hierarchical agglomeration.

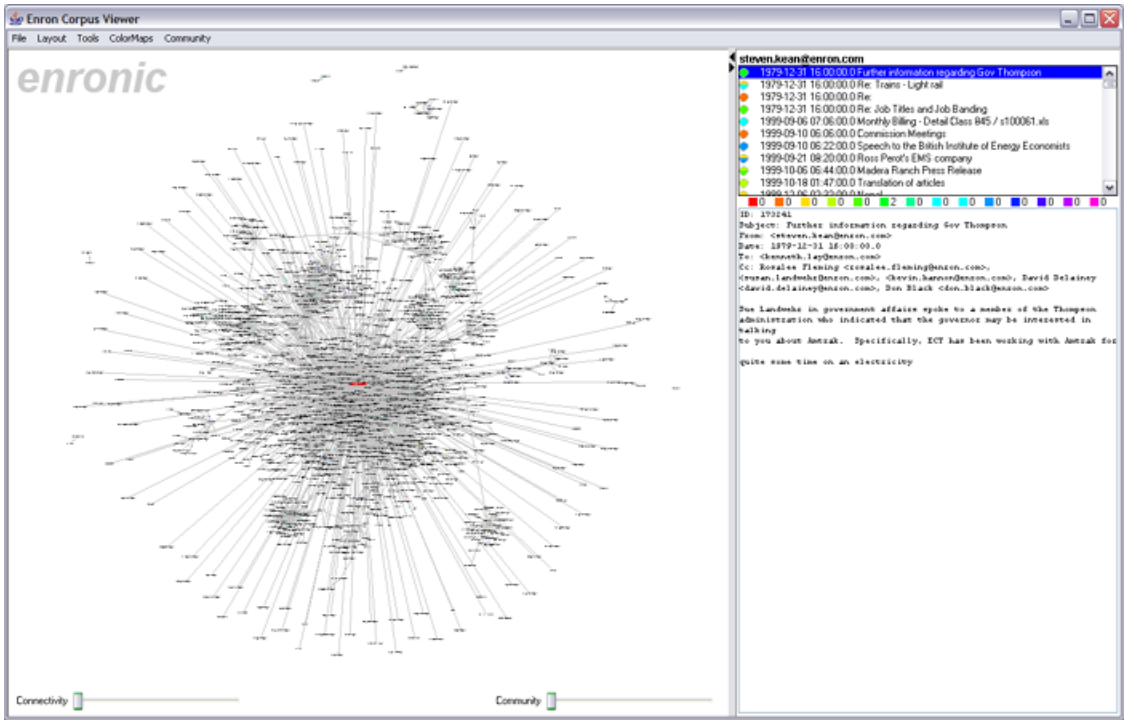

Figure 1. Enronic displays connections inferred from a set of emails, in this case the Enron archive. It provides an interactive visual representation of the network and *connectivity* (*degree centrality*) potentially indicating a person's power or influence within the network.

[42] Victoria L. Lemieux and Jason R. Baron, "Overcoming the Digital Tsunami in E-Discovery: Is Visual Analysis the Answer?" Canadian Journal of Law and Technology, 9, no. 1 (2011): 33-48.
[43] Ibid.
[44] "Exploring Enron: Visualizing AMLP Results," accessed October 9, 2012, http://hci.stanford.edu/jheer/projects/enron/v1/.

Displaying the social network implicit within an email collection is one kind of overview representation; it involves exploiting email metadata. As Lemieux and Baron[45] observe, other common forms include timelines (also based on metadata) and representations of content such as galaxy views in which similar documents are clustered together. From the perspective of a given investigation, one type of view may not be enough. And so we anticipate that systems that support people in making sense of email archives (for litigation purposes or otherwise) will incorporate multiple interactive visualizations of different types. Further, multiple, tightly coupled visualizations may be used concurrently, with users cross-referencing between, or sequentially, with different visualizations suited to different tasks in an analytic chain and at different scales of magnitude.

Sensemaking is frequently slow and painstaking. However, Visual analytic tools in conjunction hold out some promise for constructing scalable methods for making sense of large collections of documents such as emails. Just as these tools are becoming more significant for lawyers, so too we expect that they will become more significant for archivists, historians and others interested in uncovering hidden secrets that email archives may have to offer.

7. The Archivist of the Future and The Modern Digital Archive

Archivists and records managers need to confront reality in understanding that institutions are facing extraordinary pressures to adopt such automated email archiving schemes. With the encouragement of the Archivist's Directive, and in the absence of prior institutional motivation to move forward with other forms of automated or electronic recordkeeping (*e.g.,* 5015.2 compliant solutions), a present day vacuum exists: those institutions (including most of the U.S. government) that still live in a paper-based world with respect to their default, official recordkeeping system will likely leapfrog to automated capture schemes involving email archiving before they make large expenditures in deploying other forms of enterprise-wide software. The inevitable result: automated email archiving scheme will, more likely than not, also be perceived to be an institution's *state of the art records management scheme,* and it will be exceedingly difficult to convince end-users to continue any semblance of other means of recordkeeping, in the face of the knowledge that email archiving exists.

This brings us to the final, most profound challenge that current day automated email archiving schemes pose to archivists and historians: the desire on the part of institutions to save everything for the here and now, but to delete (almost) everything after a prescribed period of years (e.g., all non-permanent records). Automated email archiving has a kind of "July 4 fireworks" aspect. In the United States, July 4th – Independence Day – is a day traditionally celebrated with fireworks, which are of course both highly illuminating and highly ephemeral in nature. The profound problem posed by automated email archiving is that it has the potential to turn out to contain no "archives" of any sort – rather, it may have been conceived to date by some individuals and institutions as simply a short-term fix consisting of information that may be wholly deleted after a set term of years (without the need for any form of segregation of "permanent" record holdings). Such an archives, despite the vast amount of information contained within, would ironically leave no permanent mark. Like fireworks, petabytes or exabytes of electronic records would alight the sky for a brief period of years, and then, in terms of the institutional memory of the public sector in the 21st century, all would grow dark. The Archivist's Directive challenges the federal sector in the United States to address the need to extract "meaning" from federal

[45] Ibid.

databases by the end of this decade, by segregating permanent records for preservation in electronic form. However, all institutions confront similar issues in separating the permanent from the ephemeral.

The gauntlet being thrown down to the archival profession should be clear: records managers and archivists of all stripes demand that deployment of any such capturing schemes be accompanied by some recognition on the part of the institution that, if they are to morph into the *de facto* recordkeeping scheme for the institution, the business rules of the institution should well call for categories of "permanent" and "temporary" records being culled out and properly segregated. To the extent the IT community is empowered to build such structures, others within the organization should fairly see them for what they are, and demand front-end thinking in the procurement cycle as to the long-term records management implications of deployment. In doing so, archivists and records managers may yet be able to import "the best of" 5015.2 electronic recordkeeping into a model for automated email archiving, with greater use of auto-categorization methods, to come up with workable hybrid model.

This paper argues that we should acknowledge the obvious: end-users cannot cope with recordkeeping demands imposed by email, and thus, the end of the end user as records manager is near. But what does the implication of automated electronic archiving mean for future records managers and archivists? What role do they have to play when retention environments are "fully automated"? Records managers and archivists hold the potential to exercise enormous influence on how electronic records are managed and preserved, including ensuring that electronic archiving systems with recordkeeping functionality are sufficiently in place to allow for contemporaneous capture of permanent records in electronic forms, for immediate copying and eventual migration to the archival institution itself. This is an idea whose time has come.

In light of the reality of litigation, we believe that the "least worst" solution is for institutions to put into place automated capture solutions for email for purposes of information governance and overall risk management. We also see a need for future archivists to accept and prepare for the technological future that confronts them. This is a future where intelligent filters and automated rules-based systems replace records series and records categorization as practiced in the 20th Century; and where both front-end and back-end solutions based on notions of artificial intelligence, data mining, content and visual analytics, and the like are increasingly employed to separate wheat from chaff, the permanently valuable from the flotsam and jetsam of more ephemeral forms of records. Through such means archivists will truly make sense of the modern email archive.

Writing in The American Archivist, Richard Pierce-Moses has spoken to the demands on the archivist of the future in a similar vein:

> [A]rchivists should become as comfortable working with digital records as they are working with traditional media. Instead of pen and paper, we will work with cursor and keyboard. Instead of sorters, we will work with sorting algorithms. Rather than weeding, we will filter. With few exceptions, all archivists will need to know what we now call technical skills as the vast majority of contemporary and future records are and will be digital. Different archivists will need different technical skills . . . . No doubt some archivists will continue to specialize, but their specializations will be specific to the digital arena: databases, image and audio formats and metadata, but also user interfaces, search systems, and digital preservation.[46]

[46] Richard Pierce-Moses, "Janus in Cyberspace: Archives on the Threshhold of the Digital Era," *The American Archivist,* 70, no. 1 (2007):18. *See also* Philip C. Bantin, "Strategies for Managing Electronic Records: A New Archival Paradigm? An Affirmation of Our Archival Traditions?" (undated), http://www.indiana.edu/~libarch/ER/macpaper12.pdf.

The volume, growth, and complexity of electronic records is confronting the legal and business sectors with serious challenges, and as we have argued elsewhere, a critical juncture has been reached where 20th century ways of doing business, at least in terms of responding to discovery demands in lawsuits, no longer suffice.[47] Litigation and other forms of external access demands may be drivers of change, but they need not drive organizations off the cliff in terms of adopting reasonable solutions for the management of records both in line with short-term business needs and for history's larger purposes. Archivists and records managers, working with lawyers, IT staff, and business executives, have the opportunity to collaborate in coming up with appropriate business models that satisfactorily take into account all of these competing demands and priorities. This is not an easy road for the archivist of the future, but it is a necessary one.

Finally, we have no doubt that changing technology in this area will render some of what is said here rapidly obsolete. We trust, however, that there is more than a little value in attempting to stake out a position on these subjects, even if in so doing one must recognize that we are all engulfed inside a vortex of transformative change. The change is in our workplaces as well as in our collective cultures, with respect to what constitutes the adequate capture in modern digital archives of permanently valuable electronic records, both for business purposes as well as for the sake of history's greater illumination.[48]

[47] Paul and Baron, "Information Inflation."

[48] "So ere you find where light in darkness lies, your light grows dark by losing of your eyes."
William Shakespeare, Love's Labour's Lost