RESHAPING THE REPOSITORY: THE CHALLENGE OF EMAIL ARCHIVING

Andrea Goethals, Wendy Gogel

Harvard University Library
Office for Information Systems
90 Mt. Auburn St., Cambridge MA 02138 USA

1 ABSTRACT

Because of the historical value of email in the late 20th and 21st centuries, Harvard University Libraries began planning for an email archiving project in early 2007. A working group comprised of University archivists, curators, records managers, librarians and technologists studied the problem and recommended the undertaking of a pilot email archiving project at the University Library. This two-year pilot would implement a system for ingest, processing, preservation, and eventual end user delivery of email, in anticipation of it becoming an ongoing central service at the University after the pilot. This paper describes some of the unexpected challenges encountered during the pilot project and how they were addressed by design decisions. Key challenges included the requirement to design the system so that it could handle other types of born digital content in the future, and the effect of archiving email with sensitive data to Harvard's preservation repository, the Digital Repository Service (DRS).

2 INTRODUCTION

2.1 Value of Email

Recognizing the potential long term value of email content to Harvard's research collections, the Harvard University Library charged a working group of University archivists, curators, records managers, librarians and technologists to describe the challenges of collecting, managing and archiving email at the University and to make recommendations for possible action. The group's March 2008 report highlighted email as an essential, yet missing part of our collections, and recommended that the University Library undertake a pilot project to build a system that would enable ingest, management, basic preservation, and also pave the way for access to email. The report emphasized the administrative, historical and legal value of email to the managers of manuscript repositories, archival programs and University records at Harvard. They also recommended that we identify critical policy and curatorial issues and address any legal or security concerns.

It is now widely recognized that email represents a slice of the late 20th and early 21st centuries (so far) that will be significant to historical research in the future. For our curators[1], collecting email represents a continuation of their traditional collecting in the categories of organizational records and personal papers. Since these records do not directly replace any single genre of analog content[2], their importance to future research only begins with their function as correspondence. We now appreciate email as a complex communications package that may contain unique primary source material; often serves as the document of record for business activities, decisions and outcomes; and is critical to the preservation of recent scholarly communications. The package includes the headers, message bodies, and attachments.

2.2 The Pilot Begins

For the pilot project, we were given funding for one developer for two years. We modeled our pilot project on the Libraries' successful first born-digital project, which resulted in the establishment of a central service at the University for archiving web resources - the Web Archiving Collection Service (WAX). WAX also started as a two-year pilot project, and entailed building a system to collect, process and archive resources to the University's preservation repository - the Digital Repository Service (DRS). Like WAX, the email archiving project would be managed and developed by the Library's Office for Information Systems (OIS), and

[1] In the rest of this paper the term curator is used to refer to any Harvard collection manager including archivists, librarians, museum and special collections curators and records managers.

[2] Note that the author of [7] and [8] changed positions where [7] describes email as an equivalent to correspondence and [8] notes that it has no parallel in the analog world. Our thinking during the pilot evolved along the same path.

would involve curators from throughout the University. Staff from three University repositories - the University Archives, Schlesinger Library, and Countway Library would partner with OIS to work closely on functional requirements and to supply email collections for testing. To address the legal and security considerations, we would consult the University's Office of the General Council and the University's Technology Security Officer.

Early on, the team that was charged with implementing the pilot recognized that the challenges we were confronting from transferring data of unknown formats, through identifying and securing sensitive data, to providing authority control to manage the variations on people's names, email addresses and institutional affiliations applied also to the broader, pressing need at the University to manage all born digital content. The curators assured us that all of these issues were not new to the field, but that they simply needed new tools and work flows to manage collections that are increasingly composed of a hybrid of analog and digital content. We pledged to use an architecture that would be flexible enough to expand to other oncoming born digital content. Although the focus of the pilot project is on content that has been selected for its long term value and therefore requires deposit to our preservation repository, we envision that in the future, the central infrastructure will also need to support the temporary storage of email and other born digital content as part of the University's records management schedule.

Notably our charge from the working group did not include delivery to end users as a requirement for the pilot. To support the research and teaching missions of the University, we will eventually need to provide a user interface for online delivery of email to end users. However, it was recognized from the beginning that access issues would be too complex to address in the pilot time frame and therefore would need to be addressed in the future, after the important first steps of collecting and preserving the email. However, the pilot will need to provide a mechanism for curators to provide mediated access to the email collections for researchers and for legal discovery. In anticipation of future end user delivery, we are defining the requirements for rights management that would enable automated access restrictions to a larger audience, and would continue to support curator-mediated access to the collections.

2.3 Nature of Email

Because of the pervasive use of email, at first glance the special challenges it poses for preservation are easily overlooked. In the course of conducting this pilot, four primary challenges due to the nature of email were identified: the diversity of mail client formats, the overly-flexible structure and composition of email messages, the tendency for email to contain sensitive information, and the volume of messages typically contained in individual email accounts. Secondary challenges included the tendency for email to have viruses or spam content, and the presence of duplicate attachments within email accounts and collections.

Although the format of exchanging email messages has been formally standardized through the RFC mechanism[3], the format for storing email messages has not been standardized. The storage format, including the directory structure, packaging format and location of attachments, is left up to the developers of email clients to decide. For this reason, mail clients vary in the way that they organize content, so the particular email client software has to be taken into account when preparing email for preservation.

There are also differences among email messages that aren't related to the originating mail client. Email messages can contain message bodies in text format, HTML, or both. Technically the message bodies can be in any format, but because mail clients need to display message bodies to receiving parties, in practice, message bodies have been limited to text and HTML formats. Messages can contain attachments in any format, and can contain in-line images within HTML message bodies. Some email messages do not have message bodies - as is the case when an individual sends an email that only has attachments. All of this variation has to be taken into account when processing, indexing, displaying and packaging email for preservation.

Individuals often use the same accounts for private and business correspondence. As Clifford Lynch, Executive Director of the Coalition for Networked Information put it, "email mixes the personal and professional in an intractable hodgepodge."[4] It can be difficult to impossible to separate, especially given the quantity of email most of us have. In addition, email is considered by most a private correspondence that will never be seen by anyone other than the original receiving parties. For example, Harvard curators have acquired email in which credit card numbers have been passed, and in which private health matters have been discussed. Email is the first content likely to contain sensitive information that will be ingested into Harvard's preservation repository. As the pilot progressed, we came to the realization that the sensitive nature of email would require us to rethink and redesign our repository infrastructure.

[3] See RFC-5322 Internet Message Format, and the related MIME Document Series (especially RFC-2045 MIME Part One: Format of Internet Message Bodies).

[4] CNI Conversations, March 10, 2010.

3 PRIOR WORK

Whenever OIS begins a new large project, we always review the larger landscape for prior and current initiatives that can inform our work. About 10 years ago there was a burst of research and projects focused on email archiving and preservation. This work primarily came out of various city, state and national archives. One of the earliest of these projects, the DAVID project, was conducted by the Antwerp City Archives from 1999-2003. This project exposed many of the legal and privacy-related challenges of email archiving, and argued that email archiving solutions need to include clear policies and procedures as well as technical solutions. They chose XML as the long-term storage format for email and developed a simple XML schema for storing the message body and metadata about a single email [2].

Many other projects have also chosen XML for the normalization format for email [3][4][6]. The National Archives of Australia (NAA) created Xena, an open source format conversion tool that can convert email in three formats to an XML format. Some authors [7] conclude that text may also be a suitable long-term storage format for email. Other formats, such as HTML and PDF were considered by some but ruled out for various reasons, including the loss of significant characteristics of email or an incompatibility with search and index technologies.

Recently there were a couple of high-profile email archiving projects, also conducted by archives. The Collaborative Electronic Records Project (CERP)[5], conducted by the Rockefeller Archive Center and the Smithsonian Institution Archives, ran from 2005-2008. The Preservation of Electronic Mail Collaboration Initiative (EMCAP)[6] was conducted by North Carolina State Archives, Pennsylvania State Archives, and the Kentucky Department of Libraries and Archives. The CERP and EMCAP projects wrote guidance on transferring and formatting email, software for acquiring and processing email, and they collaborated on an XML schema designed to hold email for an account.

In addition to the DAVID, NAA and CERP/EMCAP schemas, there have been other efforts to develop XML schemas for email for general use [1] [5] [10]. In the early phases of the pilot we analyzed each of these schemas. We have preliminarily chosen to use the CERP/EMCAP schema, because we think it strikes the right balance between fully supporting the complexities of email headers and structure with a welcome lack of manipulation of the message bodies and attachments. Unlike most of the other schemas, it uses generic <Header> elements to store the names and values of the message headers. The advantage of this approach is that it can accommodate unanticipated headers, for example custom headers added by client systems, or those that will be added to future revisions of the email RFCs. It can support multiple message bodies per email, including HTML, and pointers to externally-stored attachments. They also have a separate schema for wrapping base64-encoded attachments, however we will likely decode attachments and store them in their original formats. While the CERP/EMCAP schema is designed to contain all the email messages for an account, we anticipate that it will work equally well at storing a single email message, which is how we intend to use it.

[5] See <http://siarchives.si.edu/cerp/index.htm>

[6] See <http://www.records.ncdcr.gov/emailpreservation/>

4 KEY REQUIREMENTS

To begin gathering functional requirements from our curatorial partners, we walked through several potential work flows with them. The scenarios covered the likely life cycle of email including the activities of email creators, data transfer to us, processing by the curators and then preservation in our Digital Repository Service (DRS). For the pilot project, we knew that we could not control or automate every step of the work flow and began working with the developer and other architects to refine the project scope.

A number of interesting challenges arose during this process. First, we were warned that a veritable tsunami of born-digital content was headed our way and that email would be only one of the great waves. Given the rate at which we all produce digital content, this was readily understood. Since the tsunami would include genres besides email, we are challenging ourselves to build a system that can grow and be generalized for other genres in the future. In light of the expected great wave of email, we recognized the likelihood that there would not be sufficient resources to process all of the collections at any depth. This led to the requirement to support mass transfer of content to the DRS with minimal manual processing, so that first and foremost, it would be safely and securely stored. It was determined, however, that the value of some collections would merit item-level processing of individual email messages and that this too would need to be supported. In keeping with traditional practices, curators would need to be able to return to collections that were only minimally processed and engage in more in-depth processing at a later time. This might occur because resources become available, to answer a research request, or because the value of the content has been newly assessed. This requirement - to enable processing the collection after it is transferred to the digital repository - is different than any other email archiving project we know of.

Second, we confronted new and very stringent requirements because of the potential for email to

contain sensitive data as mentioned above. Although the laws vary from country to country and even within regions (or states in the U.S.), all email archiving projects need to confront security requirements to comply with laws at multiple levels of governance as well as local security policies and practices. At Harvard, this would influence the design of the new system as well as have a profound impact on our existing infrastructure. Our email collections will likely include data that is defined by Harvard's enterprise security policy as High Risk Confidential Information (HRCI) and protected by Massachusetts State Law regarding personal information (201 CMR 17.00). Both are meant to safeguard personal information against unauthorized access or misuse and they generally cover a person's name in combination with identification numbers (such as U.S. Social Security or state driver's license numbers) or financial account information. In addition, some data will be protected by United States federal laws such as the Family Educational Rights and Privacy Act (FERPA), and the Health Insurance Portability and Accountability Act (HIPAA).

In consulting the University's information technology security experts, we discovered that email would need to be encrypted any time it was transported over a network or stored on portable media such as tape. Any applications accessing the content would need to be on the University's more secure private network, not the public network used in our existing infrastructure. Unfortunately, our DRS architecture did not comply with these security requirements. We needed to re-architect our repository to be able to accept, manage and preserve the email content. We also learned from the curators that, because of the sensitive nature of some of the content, only authorized people within the specific Harvard unit that stewards the collection would be able to view the contents.

A third challenge, reflecting the current collecting practices of our curators, is that email will represent both Harvard and non-Harvard content and may be closed or open-ended collections. The first email content contributed to the pilot project will be new content for existing analog collections. Email will be collected for noted figures in academia, science, politics and the arts (some of whom are faculty) and for institutions and organizations in areas where the University already collects "papers." This content defines the requirement to accept email from multiple mail servers, both internal and external to Harvard, from multiple types and versions of email clients, and to accept content from active as well as inactive accounts. In at least one case, the curator collected old email on a hard drive and has since negotiated with the creator to receive email on an ongoing basis.

5 DESIGN

Out of our review of prior work and the curatorial and security requirements we began to design an email archiving system that could integrate with our existing central infrastructure for authentication, authorization, persistent naming, discovery, preservation, and management. In past projects, we were accustomed to making small or no alterations to our existing infrastructure to accommodate new types of content. We envisioned adding a front end application named EASi (Email Archiving System Interface) to our infrastructure that would be used to prepare and push email into the DRS for preservation. Initially this seemed even simpler than WAX, which also acts as a specialized ingest system for the DRS, because WAX includes a complex crawler system.

After learning about the tsunami of born-digital content heading our way, we re-envisioned EASi as a front-end that could eventually accept whole hard drives of mixed content for processing and archiving to the DRS. Email would just be the first genre supported by EASi. It would now stand for the *Electronic* Archiving System Interface—not the *Email* Archiving System Interface. The EASi software developer is designing it so that it could be extended to other genres of content, and in a modular way so that the processing tools could be reused in the DRS management application, which would allow curators to continue to process the email, even after it is stored in the DRS for preservation.

Because currently there isn't a central server that we can pull email from, we are using a push model to get email into EAS. The overall data flow begins with curators transferring email to a central storage location at OIS via sftp, where an EAS process will pick it up and import it into the system. The curators will then be able to process the email using the EASi web-based interface, which will allow them to search, browse and read the email and attachments. They will be able to organize the email into collections, add rights and access restriction metadata, associate email addresses with people and organizations, delete email and/or attachments, and select content to send to the DRS. For this selected content, an EAS process will automatically prepare, package, transfer and load it into the DRS. After the content is in the DRS, curators will be able to continue to manage the email along with all their other DRS content (images, audio, etc.) using the web-based DRS management interface. As requested by the collection managers, this work flow is designed to support multiple levels of processing. Curators will have the option to do minimal processing up-front before pushing it into the DRS, knowing that they will be able to do further processing of the content later using the DRS management interface. Alternatively the system will also support more in-depth processing before pushing it

into the DRS, if the content warrants the extra up-front effort.

When email is imported into EAS, the content is put through a number of automated processes. The first process converts the email to an RFC-282 Internet Message Format [9] using Emailchemy [11]. Table 1 lists the mail formats that EAS will support in the pilot phase because they are supported by Emailchemy. All the mail clients used by the pilot collections are supported by this list except for one obsolete DOS-based client called cc:Mail. We are still investigating whether we can support this client using other tools. After format normalization the content is virus-checked, scanned for some forms of high-risk confidential information (initially just credit card and social security numbers), and scanned for spam. Finally the email content is parsed, metadata is extracted, and the metadata and content are indexed.

Email client software name and version	
AppleMail *	Outlook Express for Windows 4-6
AOCE *	Outlook Express for Mac (Database file) 5
AOL for Windows *	Outlook Express for Mac (Messages file) 5
Entourage *	Outlook for Windows
Eudora for Mac *	PowerTalk *
Eudora for Windows *	QuickMail Pro for Mac *
Mac OS X Mail 1-4	QuickMail Pro for Windows *
Mailman 2	Thunderbird
Outlook Express for Mac 4	Yahoo
Outlook Express for Unix 4	

Table 1. Formats that will be supported by EAS import

In response to the security requirements mentioned earlier, OIS system administrators came up with two options for redesigning the DRS architecture. Essentially the first option was to treat all DRS content as having sensitive data; the other option was to segregate the content coming through EASi from all the other DRS content.

5.1 Option 1: Integrated Content

In this option the entire DRS storage system would be moved to the more secure, more expensive private network. The advantage would be that we could continue to use the existing tape and disk copy infrastructure. On the negative side, we would not be able to use NFS to access the DRS files for the delivery and management applications anymore because it would be a security hole. It would need to be replaced with an ssh filesystem, which isn't known to scale at this time. In addition, the DRS management applications would need to be altered to use HTTPS connections and all DRS curators would need to access them using purchased Harvard VPN clients, even if they didn't use EASi. Because the tape backups of the EASi content would need to be encrypted we would have to purchase a separate tape library along with a SUN-encrypted backup service. Lastly, there was the concern that the front end delivery systems, which would remain on the public network, could be used to break into the back end secure system. Clearly this option had a lot of disadvantages.

5.2 Option 2: Segregated Content

In this option we would have two separate DRS storage systems – one for the content that entered through EASi, and the other for the rest of the DRS content. We would put the EASi applications and storage system on the more secure, more expensive private network, and leave the rest of the DRS on the existing network. One disadvantage is that we would need to replicate many of our DRS applications on the two different networks. We would need to replicate the DRS ingest applications that EASi uses to package and load the content into the DRS, and the DRS management application. The secured instance of the management application would need to be accessed using a Harvard VPN client. Although the content would be segregated, we could make it appear integrated from the curators' perspective because the DRS management application would be able to access both sets of content, allowing them to search and manage any of their DRS content together in the same interface. Although this option didn't allow us to leverage our existing architecture to the extent we would have liked, this seemed the better of the two options, so we proceeded to implement these changes.

6 FUTURE WORK

Although our current plan is to segregate the content coming through EASi from all the other DRS content, in the future we are optimistic that we will be able to reintegrate them. OIS system administrators are monitoring upcoming storage solutions that would allow us to have a more integrated solution when we do our next large storage migration, expected to take place in a few years.

The pilot project did not include within its scope delivery of the email content to end users—this will need to be undertaken as a separate post-pilot project. Prior to having a delivery service in place, curators and archivists will be able to access copies stored in the DRS for themselves through the DRS management application. This will allow them to provide mediated access to the email content for researchers, or if needed, for legal discovery. A delivery service for the email will entail more than the technical work of developing the delivery service application. It will also require expanded rights management metadata in the DRS, an overall strategy for collecting email at the University, and policies governing the range of activities from collection through delivery.

As we considered each of our key challenges and addressed them within the limitations of the pilot project, inevitably we thought about what we could do given enough resources. In response to the born digital tsunami expected by the curators, we envision developing an environment where curators could appraise and process incoming content in a temporary holding area until a decision about its disposition can be made. When it is determined that the content will be accessioned for long term preservation storage and access, it would be transferred to the DRS and described in one of the public catalogs. We would make another repository available for transitory content that needs to be held for a limited amount of time according to a records management schedule or for specific legal reasons. We also envision building a centralized vocabulary registry that could be used by all of the metadata services in our infrastructure to help curators with authority control of terms including the various versions of people and institutional names and email addresses found in email.

7 REFERENCES

[1] Borden, J. *An RDF XML mapping of RFC 822/MIME*, Web site, The Open Healthcare Group. 2001. <http://www.openhealth.org/xmtp/>

[2] Boudrez, F. & Van den Eynde, S. *Archiving e-mail*. DAVID project, Stadsarchief Stad Antwerpen, 2002.

[3] Dutch National Archives. *E-mail-XML Demonstrator: Technical description*, Testbed Digitale Bewaring, 2002.

[4] Green, M., Soy, S., Gunn, S. & Galloway, P. "Coming to TERM: Designing the Texas Email Repository Model", *D-Lib Magazine*, *8*(9), 2002.

[5] Klyne, G. *An XML format for mail and other messages*, Web site, 2003. <http://www.ninebynine.org/IETF/Messaging/draft-klyne-message-xml-00.txt>.

[6] Moore, R., Baru, C., Rajasekar, A., Ludaescher, B., Marciano, R., Wan, M/, Schroeder, W., & Gupta, A. "Collection-Based Persistent Digital Archives - Part 2", *D-Lib Magazine*, *6*(4), 2000.

[7] Pennock, M. *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) , 2006.

[8] Pennock, M. *Managing and Preserving E-mails*. Digital Curation Centre, UKOLN, 2006.

[9] Resnick, P. (ed). *RFC-2822 Internet Message Format*, The Internet Society, 2001.

[10] Warden, P. *An XML format for mail and other messages*, Web site, 2003. <http://www.ninebynine.org/IETF/Messaging/draft-klyne-message-xml-00.txt>

[11] Weird Kid Software. *Emailchemy - Convert, Export, Import, Migrate, Manage and Archive all your Email*, Web site, 2010. <http://www.weirdkid.com/products/emailchemy/>