DOI: 10.2352/issn.2168-3204.2016.1.0.70

Archiving Email: Relevant Business Models and Drivers of Preservation

Kristen C. Ratanatharathorn and Susanne Pichler; The Andrew W. Mellon Foundation; New York, NY

Abstract

Email is a type of born-digital record, and its preservation poses a variety of technical and philosophical challenges. This paper reports on an effort to scan the field of email preservation projects that have been initiated among nonprofit and public sector entities in the United States over the past ten years. We hope that our analysis of the challenges faced and successes achieved by these projects will help to identify opportunities for further work in this critical sub-field of digital preservation.

Introduction

An email is a digital record that can be packaged in various formats, may include conversation threads, attachments, and sensitive information, and often bridges the personal and professional lives of individuals within a single online interface, often known as a mailbox or inbox. [17] The use of this complicated form of digital correspondence is large and growing rapidly. In the workplace, a 2012 report by the McKinsey Global Institute estimated that knowledge professionals spend 28 percent of the week managing email. [4] What does this amount of activity mean for archivists?

Libraries, colleges and universities, government archives, and other cultural institutions are confronted with the challenge of capturing and preserving email in order to meet scholarly, historical, public interest, and legal requirements. Developing a business model to support email preservation strategies requires such organizations to balance the costs of meeting these requirements with benefits that are consistent with the missions of their own institutions.

This paper reports on an effort to scan the field of email preservation projects that have been initiated among nonprofit and public sector entities in the United States over the past ten years and to discern the lessons that can be learned from those efforts. Among the challenges of archiving and preserving email collections are building skilled and collaborative teams, developing technology solutions, managing collections at scale, and addressing privacy and legal concerns. Grantmaking organizations are able to play a role in some of these areas, but the long-term success of email archiving initiatives likely depends on aligning the project with the mission of the host and/or partner organizations while garnering support from institutional leadership and building a community of practice.

Methodology

We reviewed summary reports, articles, conference proceedings, and other documentation related to four completed email preservation projects: the Collaborative Electronic Records Project (CERP) led by the Rockefeller Archive Center and the Smithsonian Institution Archives; the Email Collection and Preservation (EMCAP) project at the State Archives of North Carolina, the Kentucky Department for Libraries and Archives, and the Pennsylvania State Archives; the MeMail email preservation project at the University of Michigan; and the Persistent Digital Archives and Library System (PeDALS) project led by the Arizona State Library, Archives and Public Records in collaboration with state archives in Alabama, Florida, New Mexico, New York, South Carolina and Wisconsin. In this study, we also included documentation related to two ongoing email archiving efforts: Stanford University's email Processing, Appraisal, Discovery, and Delivery project (ePADD) and Harvard University's Electronic Archiving System (EAS) in order to compare their approaches with those of the completed initiatives. We are grateful to have had helpful conversations with leaders in the field about these projects and digital preservation throughout this initial research phase. Project websites are listed in Table 1.

Table 1. Web presence of projects included in this study

Project	Website
CERP	http://siarchives.si.edu/cerp/
EAS	http://hul.harvard.edu/ois/systems/eas/
EMCAP	http://www.history.ncdcr.gov/SHRAB/ar/emailpreservation/default.htm
ePADD	https://library.stanford.edu/projects/epadd
MeMail	Reports on the University of Michigan MeMail project are available as Society of American Archivists Campus Case Studies at: http://files.archivists.org/pubs/CampusCaseStudies/CASE-14-FINAL.pdf http://files.archivists.org/pubs/CampusCaseStudies/CASE-15-FINAL.pdf
PeDALS	http://www.digitalpreservation.gov/partners/states_az.html

Findings

During the course of our research, we noted key differences in the types of collections involved in each of the six projects, including library special collections, government archives, and college and university institutional archives. Table 2 provides a high-level taxonomy of the types of collections, email records, and business purposes for preservation that emerged from this study.

DOI: 10.2352/issn.2168-3204.2016.1.0.70

Table 2. Drivers of Engagement by Collection Type

Collection Type	Primary Email Records	Business Purpose(s)
Special Collections / Research Library	Personal and organizational records from diverse sources	Scholarly access and use, Historical/cultural mission
State or National Archives	Administrative records, Correspondence of key personnel	Legal compliance, Historical/cultural mission, Public interest, Scholarly access and use
Institutional Archives	Correspondence of key personnel, Administrative records	Historical/cultural mission, Legal compliance

Table 2 is not meant to be comprehensive or definitive, but rather is presented as a descriptive tool to summarize how the organizations that we reviewed justified their email preservation efforts. For example, a special collections library that is concerned with accessioning a collection of email correspondence will likely make a different business case for preservation of and access to the materials than a university's administration would make for archiving email records of key personnel. A single organization may also be interested in email archiving for multiple purposes if they hold a variety of collections.

Relevant Business Models

Depending on the needs of an organization or project, email archiving tools may include a preservation function for long-term storage, an archival processing module, and/or an access component to facilitate use. Each function has associated costs, including storage for preservation, selection and accession of records, curation and redaction of sensitive information, and providing delivery mechanisms for access. As with other preservation initiatives, there are also potential costs associated with *not* preserving, [21] including the legal ramifications of missing emails and the reputational risks faced by cultural heritage organizations if there is a perceived gap in the historical record.

Mission-oriented, not-for-profit organizations are often highly sensitive to legal and reputational risks, and these sensitivities may have led some to initiate an email preservation effort. For example, the Bentley Historical Library at the University of Michigan was interested in email records as both a collecting institution and as the official archives of the university – motivations that were shared by the EAS project at Harvard. It is worth noting that as a public institution, Michigan is subject to statutes such as the Freedom of Information Act, and as a private institution, Harvard would not be subject to the same regulations.

Email preservation also fits squarely in the archival missions of the Smithsonian Institution Archives and the Rockefeller Archive Center, where the CERP project originated, as well the state archives involved in the PeDALS and EMCAP projects that collect and preserve records related to understanding state history and government. Finally, Stanford's ongoing ePADD project focuses on the processing, discovery, and delivery of archival email collections, which aligns with the mission of the special collections and university archives to acquire, preserve, and provide access to material. [20]

Although archiving email is an activity that supports the missions of the organizations in this study, the short-term and long-term costs of preservation are not trivial. This set of projects engaged a variety of funding sources, including grants, institutional funds from the host organization(s), and community or membership support, all of which are typically invoked to help develop and sustain digital resources. [12]

External / Philanthropic support

Grant funding is a good fit for projects that need an influx of support for an early stage or discrete phase of work. Five of the six projects in this study utilized philanthropic funds for a set of their activities. Michigan received a grant for the MeMail project from the Mellon Foundation in December 2009, which facilitated collaboration between the library and the information technology office to focus on email preservation. The Arizona State Library, Archives and Public Records had been experimenting with solutions for handling digital material in the archives when the Library of Congress National Digital Information Infrastructure and Preservation Program (NDIIPP) requested proposals for their Preserving State Government Information Initiative. In collaboration with its partners, Arizona was awarded its first grant for the the PeDALS project in 2008. This support was supplemented with further grant funding in 2009. Another collaboration among state archives, EMCAP, was funded through a grant from the National Historical Publications and Records Commission (NHPRC) to support the 2008–2009 joint project of the State Archives of North Carolina, the Kentucky Department for Libraries and Archives, and the Pennsylvania State Archives. CERP, the 2005–2008 joint Rockefeller Archive Center and Smithsonian Institution Archives project, was funded partly by the Rockefeller Foundation. The ePADD project, under development at Stanford University, received funding for Phase I (2013–2015) from the NHPRC as well as from a Stanford University Libraries grant. Phase II of ePADD development, which began in late 2015, has received funding from the Institute of Museum and Library Services. The development of Harvard University's EAS, while not externally funded, was supported through an internal grant program.

Internal / Institutional funding

Internal funding has also played a role in this set of projects, primarily to continue email archiving activities after grant funding expired. In order to receive sustained institutional support, project leaders must not only show how the initiative aligns with the organization's mission (as these efforts do), but also be able to clearly articulate a plan for delivering value over time. [12] Michigan was able to do this in an unanticipated way as part of the MeMail project. When the project team asked university personnel to select important emails for archiving, they realized that faculty and staff were not accustomed to cataloging email for preservation, and it was difficult to incorporate this practice in their workflow. Michigan was able to translate these results into an action plan by embarking on a university-wide revision of their records management and retention policies in order to facilitate email archiving across the organization. [6]

The Smithsonian has also continued their efforts following the initial grant-funded CERP project, which, in collaboration with the EMCAP team, resulted in a software tool for archiving email at the account-level rather than individual messages. [15] The Smithsonian went on to develop DArcMail (Digital Archive Mail

System), a tool that allows for searching within archived email accounts and their associated attachments. [18] In the case of EMCAP, after the grant funded period, two of the three partner institutions continued to use the tool in modest ways. The lead institution, the State Archives of North Carolina, decided not only to purchase a commercial email archiving system but also to continue to develop and use the EMCAP tool on an as-needed basis. Kentucky also continued to support the project on a limited scale, while Pennsylvania was no longer able to participate as a result of changing staff and technology priorities. [14]

Perhaps the best example of internal funding support is the Harvard EAS project, in which the library has developed an in-house solution for email preservation. Following the initial pilot phase that facilitated collaboration from multiple university departments, EAS is now available for use and is integrated with other Harvard systems, including the Digital Repository Service for long-term preservation. [11]

Community funding models

Similar to the internal funding model, the requirements for a successful community or membership funded project include demonstrating the value of the project and of ongoing participation. In this model, the case must be made to not only the host, but to all supporting institutions. The PeDALS project was envisioned as a kind of community funded model, where each of the partners would contribute to the development of shared workflows and tools for digital preservation. Sustaining this model proved to be difficult when, after the grant ended in 2012, a number of partners left the project due to budget constraints, and the New Mexico State Records Center and Archives decided to continue their preservation work with a commercial provider. Of the remaining four partners, Arizona, Alabama, and Wisconsin continued the informal collaboration by contributing funds to support a developer, while New York remained a project observer. In the absence of dedicated funding, today the project is largely dormant. [2]

Grant funding and technical development are still ongoing for the ePADD project, but community development efforts are already underway through forums for submitting use cases and contributing code. [19] The community or membership model can be an attractive option for open source projects such as ePADD that enjoy broad support from a variety of organizations.

As evidenced in the examples above, projects in this study have employed hybrid business models to address the various aspects of their email archiving efforts. In the following section, we explore the recurring themes and lessons learned from these projects to date.

Challenges and Opportunities

In these projects, the following themes emerged as both challenges and opportunities for email archiving: collaboration, staff/expertise, technology, scalability, content creation, and legal/policy considerations.

Collaboration

Collaboration among divisions within an institution and across institutions on email preservation activities promotes many of the benefits of other collaborative projects: leveraging various skills and expertise, building commitment across departmental and institutional boundaries, and designing a general system that meets the needs of a wide range of constituents. For example, the CERP project team, collaborating across two organizations, leveraged the relative strengths of each partner to develop the technology and processes to support email preservation. The Smithsonian's expertise in technology development and electronic records coupled with the Rockefeller Archive Center's experience with a variety of donor organizations helped the project to identify and address salient issues related to both technology and policy. [1] Harvard's EAS project had input from a collaborative team of archivists, librarians, records managers, and developers. Participation from across the university's departments was critical to designing a digital preservation system that would serve the needs of all stakeholders. [9]

However, the challenges of collaboration should not be overlooked, including the difficulty of aligning competing priorities, the complications of joint governance and decision-making, and the trials of securing time and funding to support collaborative work. The state archives that collaborated as part of the PeDALS project were able to develop and share a technical workflow and set of best practices for processing digital records. Yet, as noted previously, state budget cuts in 2009 made it difficult for the partners to remain committed to collaboration in the absence of grant funding dedicated to the project. [2]

Staff / Expertise

The pilot funding that each of these projects received helped to secure staff time and hire external experts. The benefits of having knowledgeable people from the library, archives, technology, legal, and administrative staff all focused on the specific challenges of email preservation was a common refrain in project reports. However, continued support and maintenance of collaborative projects is often outside the scope of team members' daily responsibilities. Finding a way to sustain such efforts beyond the grant-funded period remains a challenge.

Technology

While the metadata native to email formats (Sender, Recipient, Date, Subject, etc.) can be helpful for cataloging, the various record formats, attachments, folders, and conversational structure of email files also pose technical challenges for preservation. Archivists must find or develop software that can address these requirements.

The cost of commercial software is often too high for nonprofit organizations, and no single software offering covers the email archiving process from end-to-end. Thus, the projects in this study used open source software, sometimes in combination with commercial or home-grown solutions. The PeDALS project team was able to make use of both commercial and open source tools, negotiating a lower rate for the Microsoft BizTalk software for managing digital collections (including email) and using the open source LOCKSS system for preservation storage. [2] The CERP project team hired consultants and collaborated with the EMCAP project to develop an open source schema to convert email messages, attachments, and metadata to XML, an open preservation format that could maintain the structure of the original email account. [1] [22]

The decision to implement open source software as part of a digital preservation system presents its own set of challenges, including a lack of formal vendor support. Still, if a group of users and developers rallies around a particular technology, there is also the opportunity for community members to sustain the open source tools for continued development and use.

Scalability

Archives can be found in a range of organizations, including local and state historical societies, secondary schools, tribal colleges, research universities, religious communities, and state archives. While this diversity of organizations contributes to a rich historical record, the proliferation of electronic content has left resource-challenged organizations struggling to preserve their digital records. Off-the-shelf solutions may be prohibitively

DOI: 10.2352/issn.2168-3204.2016.1.0.70

expensive, and the range of technical expertise needed to develop an application is not likely to be found even in relatively prosperous-yet-small organizations. While financial, human, and technical resource constraints existed among all of the institutions participating in the projects we examined, project leadership tends to come from comparatively well-resourced organizations. Fortunately for the archival community, project leaders and funders expressed a sense of responsibility toward archival practitioners across the field. One of the goals of the PeDALS project was to, "build a community of shared practice including a wide range of repositories and remove barriers to technology adoption by keeping costs low." [5] CERP developers were also eager to design a solution that could by adopted by organizations with limited staff and technology resources. All but one of the components of ePADD require no IT help to implement, and the second phase of the project includes partnerships with smaller cultural institutions, which should help the developers understand the needs of users outside of research libraries.

Another aspect of scalability, referred to earlier, involves the volume and variety of records. Approaches thus far seem to favor either folder-level or message-level processing, depending on the amount of material to be preserved and the level of processing typical to an institution. A system that could manage both would likely appeal to the widest audience.

Content creation

As much as archives shape their collections through selection policies, in principle and practice, it is records creators who drive the arrangement and content of any collection. Several project leaders stressed the importance of designing applications to harmonize with work habits and priorities of record creators. At the beginning of their project, CERP project staff interviewed record creators at donor institutions to learn about the electronic records organization and environment associated with each collection. [1] The University of Michigan's MeMail project taught archivists that it would not be practical to rely upon record creators to determine the disposition of their emails individually at the time of creation. Those working with ePADD cite as one of its advantages the ease with which donors may identify sensitive content in their inboxes and the effect that such control has on their willingness to transfer records to the archives.

Legal / Policy considerations

Whether it is derived from state legislation regarding the retention and availability of official records, statutes designed to protect personal privacy and information, organizational mission, or professional and scholarly principles and best practice, [8] policy plays an important role in email preservation projects.

Policies originating outside of the archives, such as retention policies designating electronic documents as records, likely encouraged a number of projects in state or institutional archives. At times, though, competing policies can slow progress. For example, during the PeDALS project, New York State Archives staff were diverted from the project in order to meet the requirements of freedom of information requests. [5] In the EMCAP project, Pennsylvania's participation was limited by a state policy prohibiting the use of open source software. [14] Policy also shapes the requirements of a project; for example, restrictions on sharing information may have led some archives to focus on preservation rather than access because access to the materials for research may not be immediately allowed.

In designing their systems, project participants weighed existing policy, mainly developed with analog records in mind, with practical considerations. In some cases, based on what was learned in the project development process, participants were able to influence policies that touched on archives and electronic records. Although not in effect until recently, the National Archives and Records Administration (NARA)'s Capstone approach, which seeks to address the challenges of processing today's unprecedented volume of email by selecting archival records based on role rather than content, is seen by many as a viable model for managing records at scale. [13]

Discussion

One of our goals in conducting this study was to identify opportunities for further investment. Much impressive work has been accomplished by the six projects discussed above (and others not included in this study), particularly in the development of new technology and best practices for the field. Some areas for future work might include further research on the legal landscape, scaling up successful solutions, and adjusting preservation strategies to accommodate cloud-based email providers (such as Gmail).

According to the Blue Ribbon Task Force on Sustainable Digital Preservation and Access, one of the best ways to stimulate support for digital preservation activities is to demonstrate the cases for use of digital materials that have been archived. [3] Unfortunately, providing access to email collections is not always straightforward, due to legal restrictions and privacy concerns. Stanford's ePADD project is making strides in this area, by developing a tool that facilitates identification and restriction of potentially sensitive data. Further research on the legal and privacy issues associated with email archives could help nonprofit organizations make the case for use, and therefore preservation, of email records.

As noted in the earlier discussion of business models, maintaining collaborative relationships and sustaining open source development can be challenging. In order to promote ongoing maintenance and use of the tools developed to date, external investment could be helpful in building a user community through grants to cover the initial costs of adopting these technologies. Underwriting business planning efforts might be another strategy that funding organizations could take to help these projects achieve long-term sustainability.

Colleges, universities, libraries, and other cultural heritage institutions are also facing new challenges related to external email applications that present both technical solutions and policy challenges for crafting and implementing organizational email strategies. For example, Google Apps for Education is now widely used across the higher education sector. According to a Google blog post from October 2015, most of the U.S. News and World Report's top 100 universities use Google Apps. [7] Among that group is the University of Michigan, which implemented Google's Gmail, Calendar, and Documents features in the fall of 2011. [10] A January 2016 survey report on the use of Google Apps by academic librarians supports this trend. Of the 89 responding libraries, over 50% reported using Gmail "very often." [16] Further research is necessary to determine how nonprofit organizations might adjust their email preservation strategies to accommodate the growing prevalence of cloud-based email applications.

Acknowledgements

We are grateful for conversations with the following people about email archiving and digital preservation: Glynn Edwards (Stanford University), Riccardo Ferrante (Smithsonian Institution Archives), Bonnie Gordon (Rockefeller Archive Center), Virginia

Hunt (Harvard University), Skip Kendall (Harvard University), Aprille McKay (University of Michigan), Christopher Prom (University of Illinois at Urbana-Champaign), Brian Schnackel (Arizona State Library, Archives and Public Records), and Michael Shallcross (University of Michigan). We would also like to thank our colleagues Helen Cullyer, Tasha Garcia, Ellen Nasto, and Donald Waters for their helpful comments.

Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of The Andrew W. Mellon Foundation.

References

[1] Nancy Adgent and Lynda Schmitz Fuhrig, "The Collaborative Electronic Records Project Summary," Sleepy Hollow, NY and Washington DC: The Collaborative Electronic Records Project, 2009 (http://siarchives.si.edu/cerp/CERP_project_summary_122008_CC.pdf retrieved on 1/31/2016).

[2] Arizona State Library, Archives and Public Records, "Persistent Digital Archives and Library System: Final Project Report to the Library of Congress" April 19, 2012 (http://digitalpreservation.gov/multimedia/documents/PeDALS_Final_Report.pdf retrieved on 1/31/2016).

[3] Blue Ribbon Task Force on Sustainable Digital Preservation and Access, "Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information," February 2010 (http://brtf.sdsc.edu/ retrieved on 2/1/2016).

[4] Michael Chui, James Manyika, Jacques Bughin, Richard Dobbs, Charles Roxburgh, Hugo Sarrazin, Geoffrey Sands and Magdalena Westergren, "The social economy: Unlocking value and productivity through social technologies," McKinsey & Company, July 2012 (http://www.mckinsey.com/insights/high_tech_telecoms_internet/the_social_economy retrieved on 1/31/2016).

[5] Christopher A. Lee, "States of Sustainability: A Review of State Projects Funded by the National Digital Information Infrastructure and Preservation Program (NDIIPP)," Chapel Hill, NC: University of North Carolina, 2012, p. 10.

[6] Aprille Cooke McKay, "Will They Populate the Boxes? Piloting a Low-Tech Method for Capturing Executive E-mail and a Workflow for Preserving It at the University of Michigan," Society of American Archivists, Campus Case Studies, May 2013 (http://files.archivists.org/pubs/CampusCaseStudies/CASE-15-FINAL.pdf retrieved on 1/31/2016).

[7] Michael de la Cruz, "Colleges and universities find new ways to work and learn with Google for Education," Google for Education, October 22, 2015 (http://googleforeducation.blogspot.com/2015/10/colleges-and-universities-find-new-ways-to-work-and-learn-with-Google-for-Education.html retrieved on 1/31/2016).

[8] Anne J. Gilliland-Swetland, "Enduring Paradigm, New Opportunities: The Value of the Archival Perspective in the Digital Environment," CLIR, February 2000 (http://www.clir.org/pubs/reports/reports/pub89/pub89.pdf retrieved on 1/31/2016).

[9] Andrea Goethals and Wendy Gogel, "Reshaping the Repository: The Challenge of Email Archiving," in International Conference on Preservation of Digital Objects (iPRES), Vienna, Austria, 2010 (http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/goethals-08.pdf retrieved on 1/31/2016).

[10] Google, "University of Michigan unifies 19 schools under a culture of collaboration with Google Apps for Education," Google Apps Education Case Studies, 2015 (http://static.googleusercontent.com/media/www.google.com/en/us/edu/case-studies/files/university-of-michigan-case-study.pdf retrieved on 1/31/2016).

[11] Harvard University, "Overview: Electronic Archiving System (EAS)" (http://hul.harvard.edu/ois/systems/eas/ retrieved on 1/31/2016).

[12] Nancy Maron, "A Guide to the Best Revenue Models and Funding Sources for your Digital Resources," Ithaka S+R, March 27, 2014 (http://sr.ithaka.org?p=22805 retrieved on 1/31/2016).

[13] National Archives, "Email Management," (https://www.archives.gov/records-mgmt/email-mgmt.html retrieved on 2/4/2016).

[14] North Carolina Department of Cultural Resources, "Final Report: March 1, 2008 – June 30, 2009," final report on the NHPRC EMCAP Tool, (http://www.history.ncdcr.gov/SHRAB/ar/emailpreservation/docs/emcap_finalreport_20100109.pdf retrieved on 1/31/2016).

[15] North Carolina State Archives and the Smithsonian Institution Archives, "E-Mail Account XML Schema Documentation," (http://www.history.ncdcr.gov/SHRAB/ar/emailpreservation/mail-account/mail-account_docs.html retrieved on 2/1/2016).

[16] Primary Research Group Inc., "Academic Librarian Use of Google and Its Apps and Features," 2016.

[17] Christopher J. Prom, "Preserving Email," Digital Preservation Coalition Technology Watch Report, 2011 (DOI: http://dx.doi.org/10.7207/twr11-01 retrieved on 1/31/2016).

[18] Lynda Schmitz Fuhrig, "Yes, We're Still Talking About Email," Smithsonian Institution Archives, August 4, 2015 (http://siarchives.si.edu/blog/yes-we're-still-talking-about-email retrieved on 1/31/2016).

[19] Stanford University, "ePADD," ePADD project website (https://library.stanford.edu/projects/epadd retrieved on 1/31/2016).

[20] Stanford University, "Special Collections & University Archives," (http://library.stanford.edu/libraries/spc/about retrieved on 1/31/2016).

[21] Tyler O. Walters and Katherine Skinner, "Economics, sustainability, and the cooperative model in digital preservation," Library Hi Tech, vol. 28, no. 2, pp. 259-272, 2010.

[22] W3C, "Extensible Markup Language (XML) 1.0 (Fifth Edition)," W3C Recommendation, November 26, 2008 (https://www.w3.org/TR/xml/ retrieved on 1/31/2016).

Author Biographies

Kristen C. Ratanatharathorn is the senior program associate in the Scholarly Communications program at The Andrew W. Mellon Foundation. She received a bachelor's degree in business from the University of North Carolina-Chapel Hill, and a dual master's degree in international and world history from Columbia University and the London School of Economics. Previously, Kristen served as an information technology consultant at Accenture.

Susanne Pichler is the librarian of The Andrew W. Mellon Foundation. She holds a bachelor's degree from Barnard College and a master's degree from the Columbia University School of Library Service. Before joining the Foundation, Susanne worked in the New York Public Library Research Libraries and for the Planned Parenthood Federation of America.

