



**CoSA-NHPRC Symposium  
Government Email in an Age of Risk:  
Preventing Information Loss  
September 15, 2017**

**Case Study 4:**

**Preservation Formats**

**Email Preservation of Incomplete Collections at the Massachusetts Archives**

*Veronica Martzahl, Electronic Records Archivist, Massachusetts Archives*

**Background**

The digital preservation program at the Massachusetts Archives (MA) has been in place since 2013. The primary impetus for the program was an impending gubernatorial transition. The MA was determined to be able to accession born-digital and digitized materials from the outgoing administration of Governor Deval Patrick.

Like many governmental entities, the Massachusetts Archives' collection policy is dictated by a records retention schedule. In Massachusetts, a six-member board known as the Records Conservation Board<sup>1</sup> oversees the creation and implementation of the Statewide Electronic Records Retention Schedule<sup>2</sup>, although the Office of the Governor has at various points in time claimed exemption from the Schedule. The Patrick Administration, however, pro-actively sought out the MA for consultation on transferring records to our custody.

The MA used a combination of the existing Schedule along with guidance from previous transfers of physical materials to determine a set of records for transfer. This proved problematic in terms of email. Most email functions as correspondence, and there is not a schedule specifically addressing correspondence that has a *Permanent* retention period. It is far more common for correspondence to be included in other, broader schedules as a component or supporting part of a larger record. For example, correspondence might be included with personnel or contract files and be bundled under the retention period for those records. This

---

<sup>1</sup> Records Conservation Board website [www.sec.state.ma.us/arc/arccrb/rcbidx.htm](http://www.sec.state.ma.us/arc/arccrb/rcbidx.htm) (Accessed July 24, 2017)

<sup>2</sup> Massachusetts Statewide Records Retention Schedule database [retweb.sec.state.ma.us/retweb/default.asp](http://retweb.sec.state.ma.us/retweb/default.asp) (Accessed July 24, 2017)

structure of including correspondence as part of larger schedules encourages pulling email out of its overall context of the email client and stresses message level appraisal and filing. As a result, the MA received only a small cache of email from one office and this cache was primarily individual messages removed from their larger context.

### **Project Methods and Scope**

Because the digital preservation program was just getting underway to be able to address digital records, the MA began reviewing options for a Trustworthy Digital Repository system. Ultimately, we chose to purchase the Preservica Standard Edition<sup>3</sup> One component of the Preservica system that was of specific interest was its ability to ingest email. While the selection of this tool did move us in the direction of converting our email messages to .eml file format during the ingest process, we felt it necessary to review other format and tool options available for the preservation of these records.

Our primary concern for the small amount of email was preservation. On the positive side, that meant that we were not overly concerned with issues of confidential, sensitive or protected information within the email messages. We were also working with largely individual messages, so issues of preserving the coherent nature of the full email account were not pressing at this time, but we remain hopeful that future changes to the Retention Schedule may allow for more complete accounts to be transferred.

When reviewing the work done around email within the archival profession to date, it is apparent the bulk of the early work focused on preservation, over and above issues of access. This has been slowly changing with the work of the Library of Virginia<sup>4</sup> and with ePADD<sup>5</sup>, but most of the tools developed prior to this looked to translate the disparate file formats created by email systems into a XML schema<sup>6</sup>. These schema would preserve the content of the messages as well as the context of the accounts. Systems also had varying means of supporting attachments to the files. (See Appendix A)

Following review of other options and tools, it was determined that, given limited technical support and expertise to install and maintain an additional system, the MA would utilize the email preservation tools within the Preservica TDR.

---

<sup>3</sup> As of June 2017 the MA is upgrading our Preservica system to the Enterprise Edition

<sup>4</sup> Kaine Email Project. Library of Virginia <http://www.virginiamemory.com/collections/kaine/> (Accessed July 24, 2017)

<sup>5</sup> ePADD Project. Stanford Libraries. <https://library.stanford.edu/projects/epadd> (Accessed July 24, 2017)

<sup>6</sup> For examples, see the Collaborative Electronic Records Project (CERP) from the Rockefeller Archive Center and the Smithsonian Institutional Archives <http://siarchives.si.edu/cerp/parserdownload.htm> (Accessed July 24, 2017) and the Preservation of Electronic Mail Collaborative Initiative (EMCAP) by the North Carolina State Archives, the Pennsylvania State Archives, and the Kentucky Department of Libraries and Archives <http://siarchives.si.edu/cerp/parserdownload.htm> (Accessed July 24, 2017)

### **Issues and Challenges**

In Preservica itself, the process of ingesting email varies from edition to edition. The Cloud Edition which is utilized by much of the company's American clientele often is used as the demonstration system for Preservica trainings, so it has been challenging to understand the process of ingesting email records into the Standard edition.

The MA's greatest challenge regarding email is the Retention Schedule. Until that can be changed to reflect the importance of the work and correspondence that is often carried out in email and given a permanent retention period, the MA will struggle to receive any significant quantities of this format. Additionally, best practice for the preservation of email focuses on preservation at the level of the email account and we are by necessity left to work with email at the message level instead. A broader discussion, beyond the scope of the MA, is needed to assess whether a capstone approach may be warranted or whether correspondence as a whole should be reclassified.

### **Lessons Learned**

Email continues to be a complex and perplexing issue within government archives. Often the technical issues are just the start of the puzzle, with issues of policy circumventing retention and preservation. The most robust technical solutions are ultimately limited if they are not supported by guidance to records creators about the importance of the records they are creating.

Additionally, email mirrors much of work in digital preservation, where initial emphasis of research is placed on preservation of records and work on access follows later. Early attempts at email preservation focused on use of XML schema for preservation of content and context while the more recent work is trying to incorporate access as well, including all of the complexities of redaction.

### **Questions remaining**

The MA has struggled with the ingest process for email in Preservica and it still remains to be seen exactly how these records will eventually be accessible for internal research by Archive's staff. We also expect to eventually have email that would be publicly accessible and it is unclear how we will address issues of redaction within the system.

### **Summary**

The MA Archives has chosen to utilize the email preservation tools available within our chosen TDR, Preservica. This has led us to prioritizing preservation of messages in .eml format and opens up additional questions for how we will ultimately support issues of redaction and access.