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

Case Study 5:
Real World Examples of Email Access Issues
Virginia's Governor Tim Kaine Email Project

Kathleen Jordan, Digital Initiatives and Web Services Manager and
Roger Christman, Senior State Records Archivist, Library of Virginia

Background

The Library of Virginia was created by the General Assembly in 1823 to organize, care for, and manage the state's growing collection of books and official records, many of which date back to the early colonial period. The agency's responsibilities have expanded over the years, as outlined in section 42.1-79 of the Virginia Public Records Act:

> The archival and records management function shall be vested in the Library of Virginia. The Library of Virginia shall be the official custodian and trustee for the Commonwealth of all public records of whatever kind, and regardless of physical form or characteristics, that are transferred to it from any agency.

In 2005, the Library accessioned its first true transfer of born-electronic gubernatorial records. Since then, we have developed policies in support of the creation, transfer, and management of this content. We understand and take seriously our responsibility to ensure secure and stable management of this material, as well as to provide open and free public access to the archival records of our government regardless of format.

Budget challenges, staff vacancies, and the absence of definitive professional best practices hindered the Library's ability to address processing and access needs for these materials quickly. However, former governor Tim Kaine's April 2011 announcement of his candidacy for the United States Senate and the potential inquiries regarding his administration's records (2006-2010) gave us the opportunity to reassess and reconsider our priorities around these records, especially the born-electronic materials.

With our senior leadership's support, a workgroup of archivists and IT staff undertook the challenge of making the Kaine administration's emails accessible to the public in time for the 2012 election cycle. Due to the sheer volume of emails (approximately 1.3 million), coupled with technical challenges and limited resources, we did not meet that deadline, but continued to move forward.

Explanation of work/project methods and scope

The Library wrestled with a number of issues during this project, but first and foremost was the question of how the processed emails should be served to the public. Given a choice of limiting access to dedicated computer terminals in the Library's reading room or allowing anyone with an internet connection to view the emails through the Library's online digital asset management system (DAM), the Library chose to provide online public access to the processed materials via our DAM, DigiTool. This seemingly simple decision had a significant impact on processing the collection, as well as our eventual understanding of public access and use.

Providing online public access to this collection committed us to a close, item-by-item processing protocol that balanced open access with the various laws that restrict access to legally protected information and the removal of any non-record material included in the original transfer.[1] Using the appropriate retention schedules, archivists reviewed every email in each email account and segregated the emails that did not qualify as public records or were otherwise restricted from public access. Processed copies of the email PST files were then passed on to the Library's information technology department for the technical phase of the project.

We chose PDF as our access format. We reasoned that users are familiar with the format, it is probably the most sustainable format that is widely used by the public, and it makes the emails look like emails. Low-cost software allowed us to convert the emails in the PST files to full-text searchable PDFs, which also included conversion and inclusion of attachments.

We automated the metadata creation as much as possible, using the information in the headers of the emails along with boilerplate administrative details to create each descriptive record. The title for each item's metadata record was sourced from the "subject" field of each email, leaving titles largely at the mercy of the person who created the first message in the string. This did not strike us as a huge problem for discoverability since we were also providing the full-text of each message, attachments included. However, that amount of searchable data comes at a

[1] For example, see Code of Virginia, 2.2-126 Disposition of official correspondence at http://law.lis.virginia.gov/vacode/title2.2/chapter1/section2.2-126/ and 42.1-78 Confidentiality safeguarded at http://law.lis.virginia.gov/vacode/title42.1/chapter7/section42.1-78/

price, and it was at this point that we confronted head-on the inherent limitations in our current digital asset management system to deliver such a large, text intensive dataset.[2]

Given that results set for a given keyword search of the Kaine Email Collection could easily number in the thousands due to the sheer volume of total records in the collection, we felt users would benefit from a variety of research tools. To facilitate use, we created the Kaine Email Project @ LVA website. From this page, researchers access the collection, search tip sheets, a collection finding aid, and related analog and digital collections.

Issues and challenges encountered in work/project

Officially launched in January 2014, the project garnered attention and positive feedback. Largely lauded as a valiant move toward open government and transparency, colleagues in the library and archives profession, members of the media, political bloggers, and open government advocates acknowledged the important step we had taken. That year the project was awarded the Open Government award from the Virginia Coalition for Open Government (VCOG) and the Council of State Archives (COSA) Rising Stars Award. However, our understanding of use of the collection and how people might access it is something that has evolved over time thanks to several purposeful efforts on our part, as well as events largely out of our control.

One of the challenges the Library faced was how to promote the collection beyond the initial press release announcement in January 2014. While the email is available online, it is within a database, and the contents are not discoverable by web search engines. This project pushed us to think about marketing the collection and making sure that we kept it on the world's radar after the initial launch. Roger Christman, the principal archivist on the project, promoted access to the collection, announcing new releases of material with timely posts of Kaine email content relevant to current issues of public interest on the Library's Out of the Box blog. The posts were promoted across the Library's Facebook and Twitter accounts. Perhaps most importantly, this created a Kaine email "digital footprint" discoverable by search engines.

At the same time, the LVA team was introduced to Gordon Cormack of the University of Waterloo and Maura Grossman, a prominent e-discovery attorney, who were looking for a large dataset for their Trec 2015 research project. TREC is the Text REtrieval Conference, sponsored by the National Institute of Standards and Technology (NIST), which supports research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies. Cormack and Grossman were part of the Total Recall Track, which evaluates Technology Assisted Review (TAR) systems. LVA allowed the research team to use the PST files of the public Kaine content to test various systems, using a continuous

[2] More detailed information on this process is provided at http://www.virginiamemory.com/collections/kaine/under-the-hood.

active learning (CAL) protocol. Essentially, during this process one or more relevant documents (either real or hypothetical, which is known as a synthetic document) are presented to a machine-learning algorithm. In the next step, the algorithm suggests the next most-likely relevant documents. The user then reviews the suggested documents and provides relevance feedback to the learning algorithm, indicating whether each suggested document is actually relevant or not. Steps 2 and 3 are repeated until very few, if any, of the suggested documents are relevant. During this very interesting experience, LVA staff learned quite a lot about the possibilities of machine-assisted review for the future of archival electronic records processing. For more on this process, see http://cormack.uwaterloo.ca/caldemo/AprMay16_EdiscoveryBulletin.pdf.

The work of Cormack and Grossman coalesced with Hillary Clinton's announcement of Tim Kaine as her running mate in the 2016 presidential race, together catapulting the collection into the limelight. Happily, we were able to respond effectively, fairly efficiently, and in ways we couldn't have foretold.

Problems solved, successes, failures, and lessons learned through work/project

The use of the collection as outlined above is largely thanks to promotion of the materials in ways that highlighted very specific components of the collections, and the availability and expertise of Roger Christman. These specific uses are probably more numerous than for most other government records collection, especially in such a close time frame.

Roger's awareness of current issues and the blog posts appeared to catch the attention of the press. After being contacted by Roger regarding the collection, David Ress of *The Daily Press* wrote a 2014 article using the Kaine email to draw parallels between that year's state budget impasse and Kaine's efforts in 2006 to avoid a state government shutdown. That same year, James A. Bacon, political blogger of *Bacon's Rebellion*, tied the 2014 opening of the Washington Metro Silver Line extension to Tysons to the collection. He lauded its potential as a boon to future researchers of the extension and "the extraordinary political maneuvering it took to make [the Metro expansion] happen" as depicted in the Kaine email collection. And a 2015 *Loudoun Times-Mirror* article on the Virginia Tech massacre of 2007 told the story through the emails of the Kaine administration.

Tom Kapsidelis, a former editor at the *Richmond Times Dispatch*, is currently working at the Library of Virginia as a Virginia Foundation for the Humanities (VFH) Research Fellow. Tom made extensive use of the Kaine email collection for his forthcoming *Higher Aim: Guns, Safety, and Healing in the Era of Mass Shootings*. Kapsidelis was aware of the Kaine email when he applied for the VFH fellowship, during which he connected with Christman who helped guide him through the records. As Tom said, "Being aware of the emails and knowing more about their organization and accompanying documents are two different things entirely."

Vivian E. Thompson, University of Virginia environmental science professor and former member of Virginia's Air Pollution Control Board (2002-2010), draws heavily on email correspondence in the Kaine collection in her book *Climate of Capitulation: An Insider's Account of State Power in a Coal Nation* (2017). Thompson's discovery of the Kaine collection relied upon the mediation of one of today's most common research assistants: Google.

In 2015, Code For Hampton Roads, a part of the Code for American Brigade, requested a chance to tackle this massive data set as part of the third annual National Day of Civic Hacking. They used an export of the data from our DAM which included xml metadata files for each object (email) and the corresponding html full-text extractions and pdfs. The hackers' goal was to devise new entry points for researching the collection, such as visualizations of topic frequency in Kaine administration email discussions, maps showing which correspondents interacted with each other the most, and a "word cloud" of the most common terms used in the set of emails available for public viewing.

Usage statistics on the collection show the impact of the presidential campaign and the research partnership with the University of Waterloo. From launch on January 14, 2014, through June 30, 2016, 10,196 (11.35 per day) items in the collection were viewed, although we do not know how many different, or unique, visitors or LVA staff this included.

In June 2016, only 158 emails were viewed (5 per day). After Hillary Clinton announced Tim Kaine as her running mate on July 22, 2016, the numbers increased dramatically:

Dates	Item Views	Average # Items per day
July 1 – July 21, 2016	807	40.35
July 22 – November 8, 2016	4,951	45.43
November 9 – July 13, 2017	14,396	58.5

Why the post-election numbers are higher is not clear, though the work of Kapsidelis and Thompson might have had some impact. However, it would probably be fair to say that this collection is the most used Virginia governor's collection ever, excepting perhaps those of Thomas Jefferson or Patrick Henry. Certainly, the amount of use in such a short time frame since release is beyond any other and can be broken down a bit more, with the following articles quoting Kaine email:

- *Politico*: July 22, 2016 (2), July 23, 2016, August 2, 2016, and August 17, 2016
- *Washington Post*: July 23, 2016
- *New York Times:* July 24, 2016
- *Bloomberg BNA (Bureau of National Affairs)*: August 15, 2016
- *New Yorker*: October 24, 2016

Politico writer Darren Samuelsohn contacted Roger Christman on several occasions about how to search the collection and for guidance on content included in the collection, which helped him understand what would be there and what would not (crash course in records management!). Conversely, none of the reporters from *The Washington Post, Bloomberg BNA,* and *The New Yorker* consulted with the LVA at all. *The New York Times* data journalist requested a data export of the collection; happily we were able to repurpose the data export sent to Code for Hampton Roads.

Our partners at the University of Waterloo also promoted their work using the Kaine email collection soon after Clinton's announcement. This included articles in the following legal journals:

- *LAW.COM*: July 27, 2016
- *ABA Journal*: July 28, 2016
- *Legal Tech News*: August 1, 2016

While the writers of these articles didn't necessarily use the collection themselves, the promotion of the collection as a dataset garnered interest among the e-discovery community. Between August 2016 and July 2017, four e-discovery or machine-learning professionals contacted the LVA for access to full exports of the publicly available Kaine email content. Again, we sent the Code for Hampton Roads data, which is now on an ftp server and is augmented with each successive release of project email. The most recent request came in July 2017.

The successful use of the Kaine email by Waterloo's continuous active learning tool led to continuing collaboration with the Library. In late 2016, the Library began tests to determine the effectiveness of e-discovery software tools on a large batch of unprocessed Kaine emails. The results have been encouraging and could lead to more efficient and quicker processing, which could have considerable impact on public availability and use.

Questions remaining, ramifications or further work or research in work/project

1. How do we balance the potential of "immediate" access to archival email collections with ensuring that we develop a solid long game?
2. How should we view return on investment with these collections? How do we manage others', and our own, expectations regarding use of huge electronic collections? Most of which, at this point, won't come near this blip with the Kaine email.
3. Conversely, does this project show the potential of modern, electronic records collections as widely known, useable resources to the public, media, researchers, etc.? And how do we harness that potential to our benefit?
4. And how does that change the way we conceive access / presentation of these collections? What are the opportunities of big data, open data, and linked data? Is there still a place for traditional archival description and item level access? How do we combine the two?

5. If TAR and CAL turn out to be valid processing tools, and materials are visible and useable faster – what's the fall out? If public availability of email records becomes the norm, how will creators react? Better records management? More or less transparency?

Summary

The Library does not claim to have a perfect solution for email access. As a matter of fact, we don't believe there is a solution at all. From this project, we take away the knowledge that email access requires flexibility and creativity. Our goal was to do the best we could with our available resources and be as responsive to changes and opportunities as possible. We were also lucky; without Tim Kaine running for VP in 2016, the collection would not have generated the use and subsequent access questions listed above.

What we didn't anticipate was our how the collection would be embraced by the e-discovery community. One of the weaknesses of the project, as we perceived it, was the discovery limitations of our DAM. However, the system made it easy to provide bulk downloads that are easy for data professionals to use, opening another avenue to access we hadn't considered at first.

The Library's partnership with the University of Waterloo could lead to the creation of a tool that would automate a large portion of email processing, leading to quicker access and the attendant questions that raises. The irony of item-level processing leading to this outcome is not lost on us. It has already made every minute spent manually reviewing those emails and wrestling them into our DAM worthwhile.