From: Processing the Unprocessable
To: Accessing the Inaccessible

Subject: TOMES, NLP, and Government Email

Discussion

1. Why is Email a problem?
2. Approaches to Email
3. TOMES
4. EAXS

Why is Email a problem?

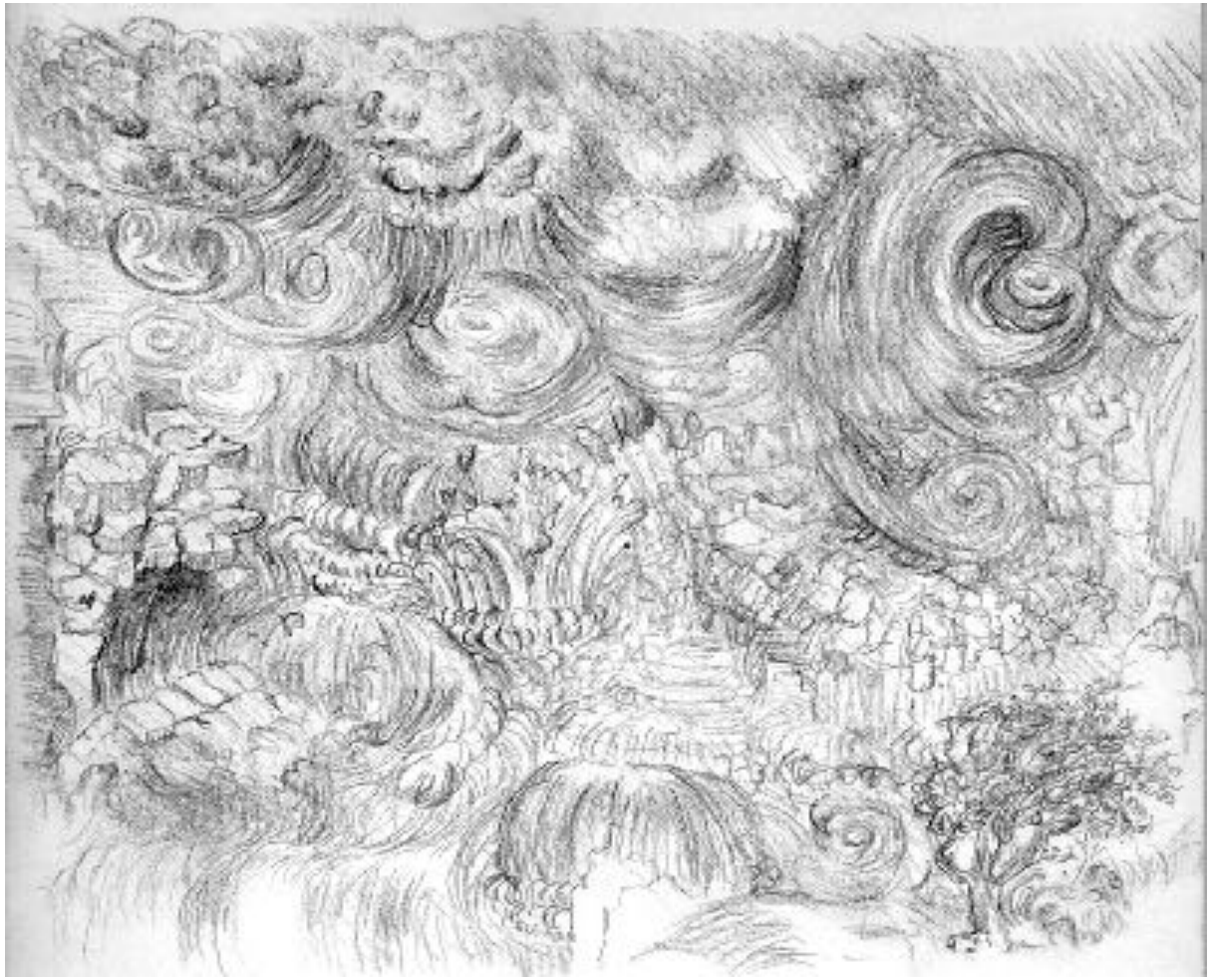
103,186,233

Why is Email a problem?

35,083,319

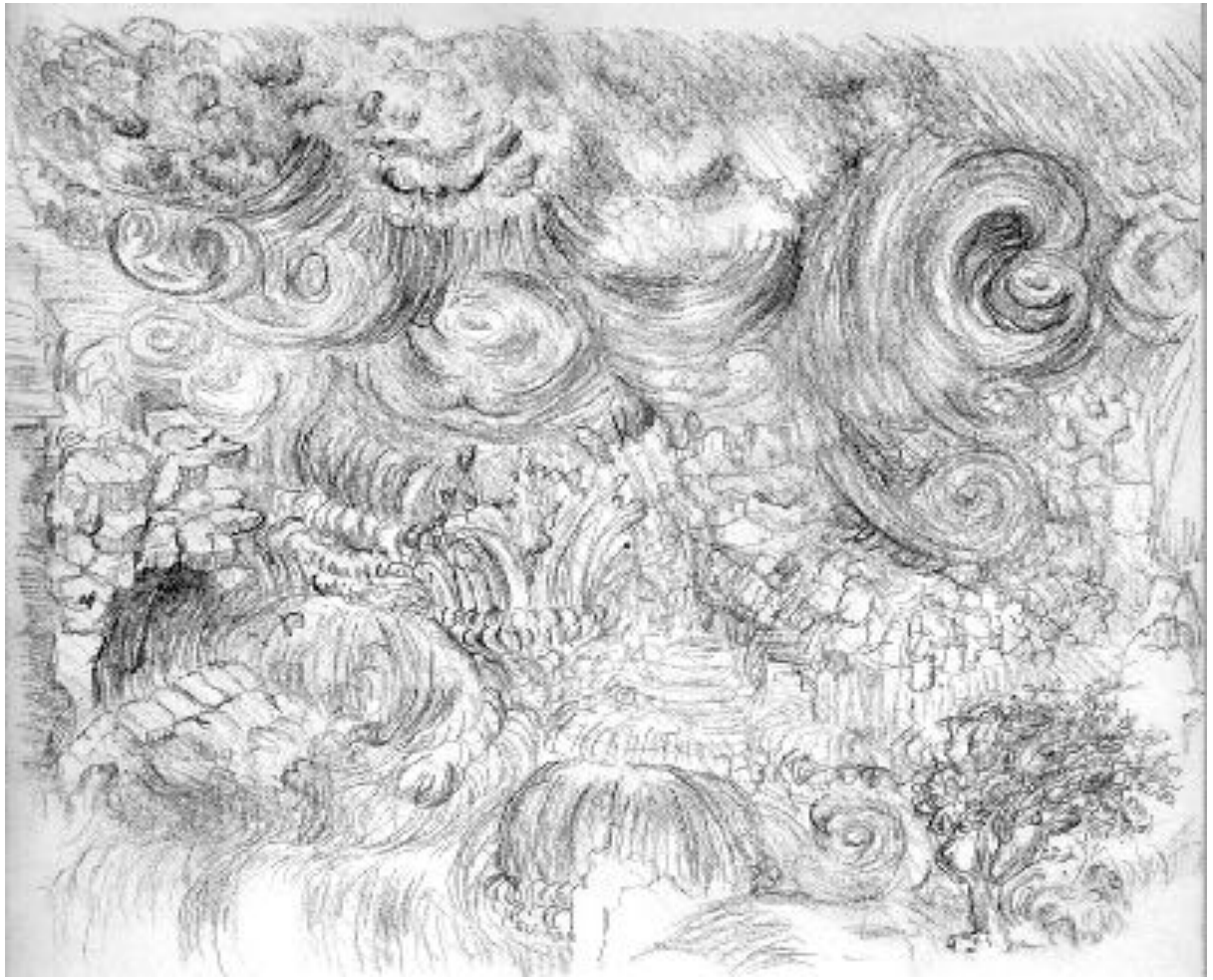
Why is Email a problem?

Those numbers should illustrate why email is the most difficult problem that archivists are facing in the 21st century.



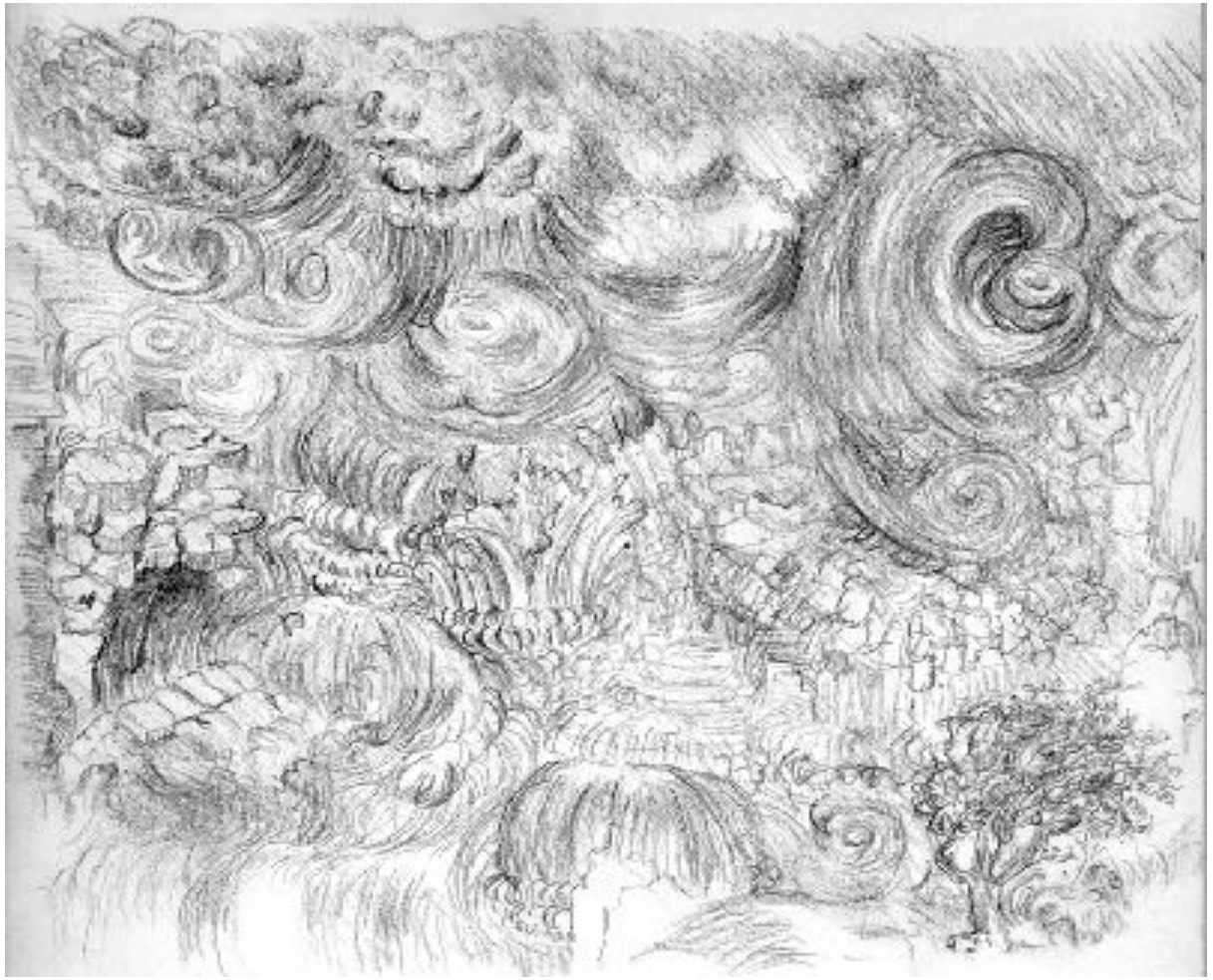
Why is Email a problem?

But, wait, that's not all …



Why is Email a problem?

Other factors beyond sheer volume make email an even trickier problem.



Why is Email a problem?

Born Digital

- Bit Rot
- Obsolete hardware
- Obsolete software
- Proprietary formats
  - *PST*

Why is Email a problem?

Legally dispositive correspondence

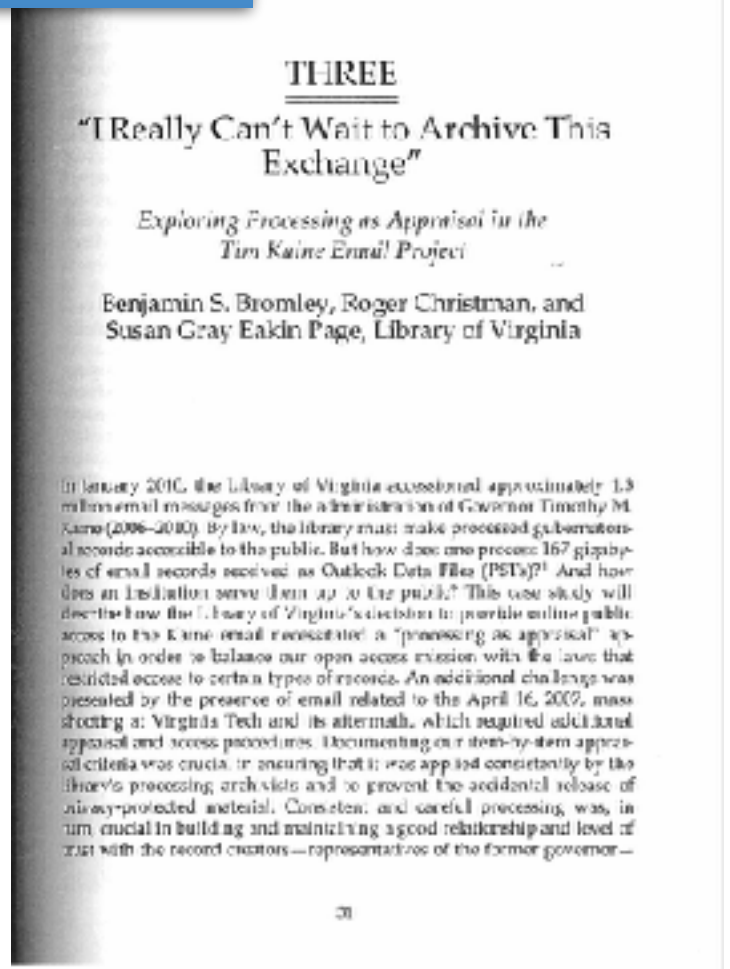
- Authenticity
  - *Chain of custody: depositor to archive*
  - *Audit of changes: to prune or not to prune?*
- Integrity of Metadata
  - *Do current tools ensure we've got everything?*

Approaches to Email

Library of Virginia

- 2010-2014
- 1.3 million emails
- 562,000 emails individually processed
  - ***Respect!***

Despite our best efforts at the time, we now realize that a different approach to records management for electronic records is required. People do not create and use electronic records in the same way that they do paper records. As of September 2014, of the nearly 562,000 emails we



THREE

"I Really Can't Wait to Archive This Exchange"

Exploring Processing as Appraisal in the Tim Kaine Email Project

Benjamin S. Bromley, Roger Christman, and Susan Gray Eakin Page, Library of Virginia

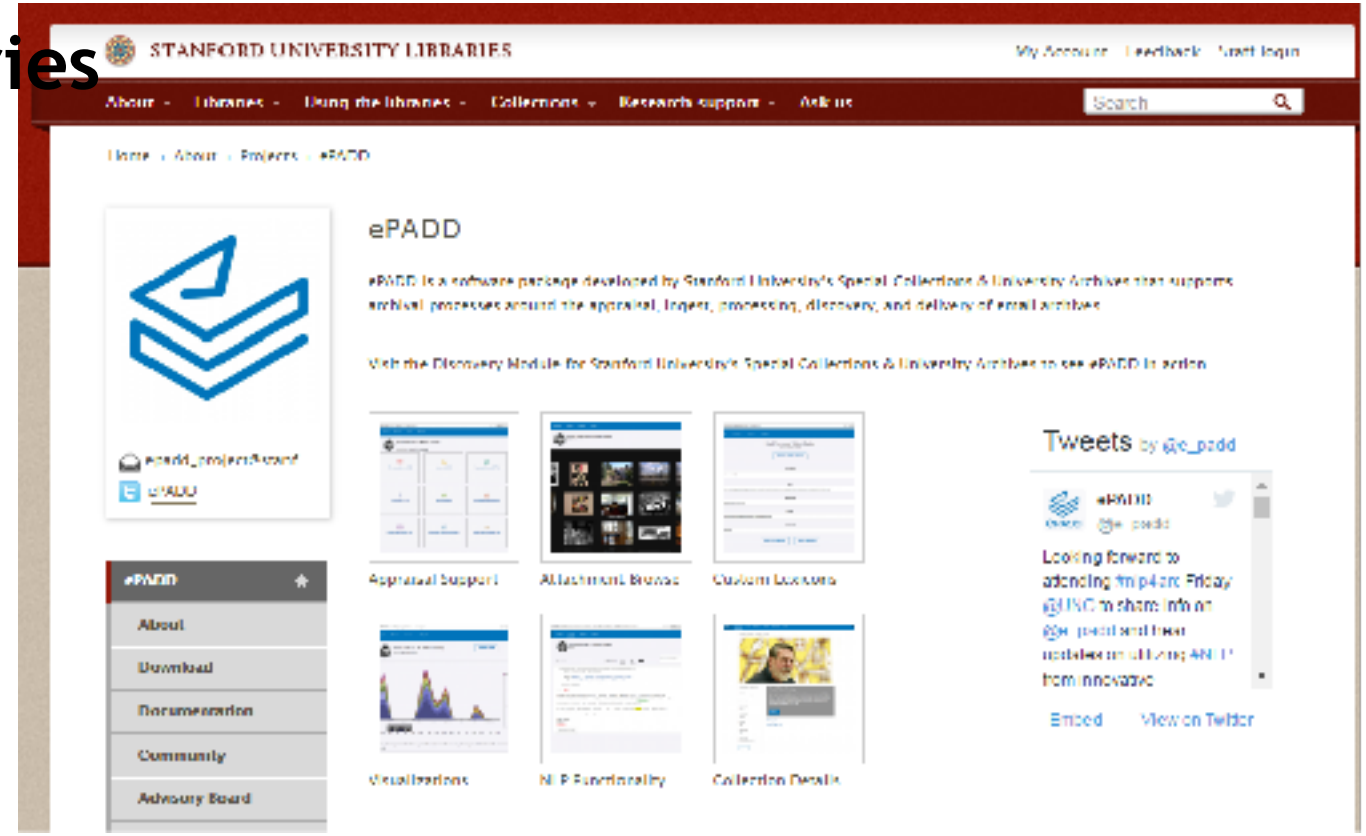
In January 2010, the Library of Virginia accessioned approximately 1.3 million email messages from the administration of Governor Timothy M. Kaine (2006–2010). By law, the library must make processed gubernatorial records accessible to the public. But how does one process 167 gigabytes of email records received as Outlook Data Files (PSTs)? And how does an institution serve them up to the public? This case study will describe how the Library of Virginia's decision to provide online public access to the Kaine email necessitated a "processing as appraisal" approach in order to balance our open access mission with the laws that restricted access to certain types of records. An additional challenge was presented by the presence of email related to the April 16, 2007, mass shooting at Virginia Tech and its aftermath, which required additional appraisal and access procedures. Documenting our item-by-item appraisal criteria was crucial in ensuring that it was applied consistently by the library's processing archivists and to prevent the accidental release of privacy-protected material. Consistent and careful processing was, in turn, crucial in building and maintaining a good relationship and level of trust with the record creators—representatives of the former governor—

31

Approaches to Email

Stanford University Libraries

- ePADD



NLP to the rescue …

TOMES

Transforming
Online
Mail
with
Embedded
Semantics

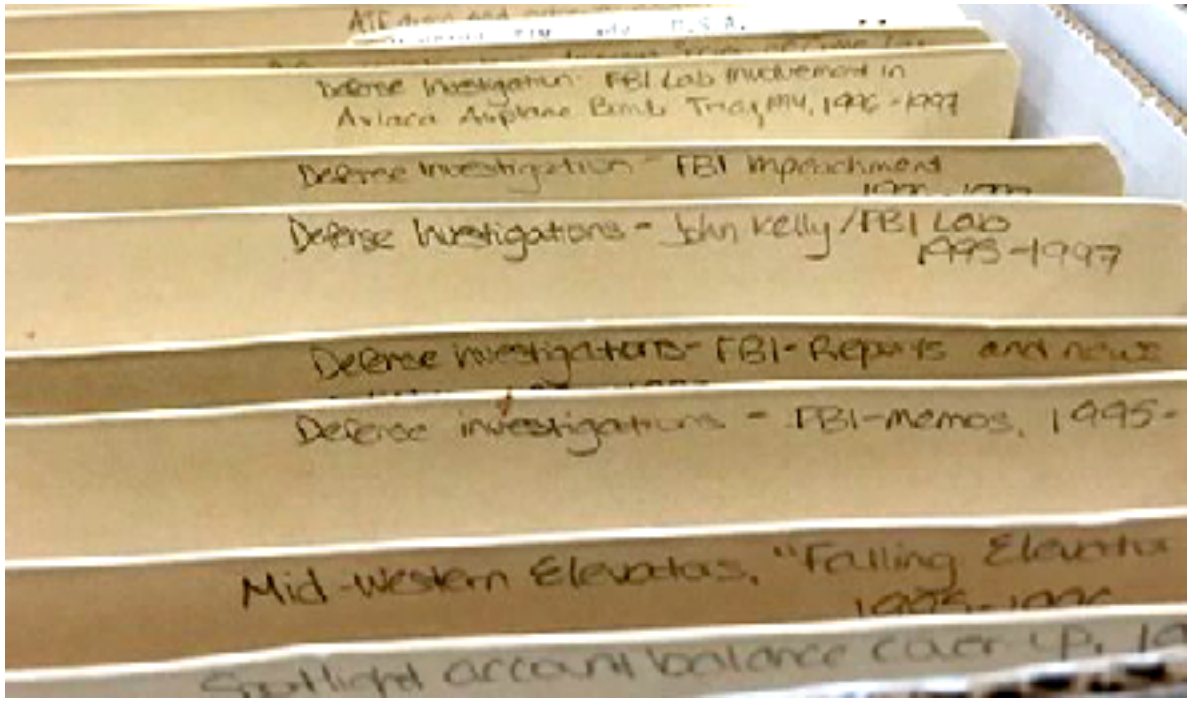
TOMES

- NHPRC State Government Electronic Records Grant
- 2015 – 2018
- Partnership between State Archives of NC, Utah State Archives, and Kansas State Historical Society
- Advisory group includes Cal Lee (UNC-Chapel Hill), Chris Prom (University of Illinois Urbana Champaign), and staff from the Library of VA

TOMES

Can 35 million records be made accessible in the traditional sense?

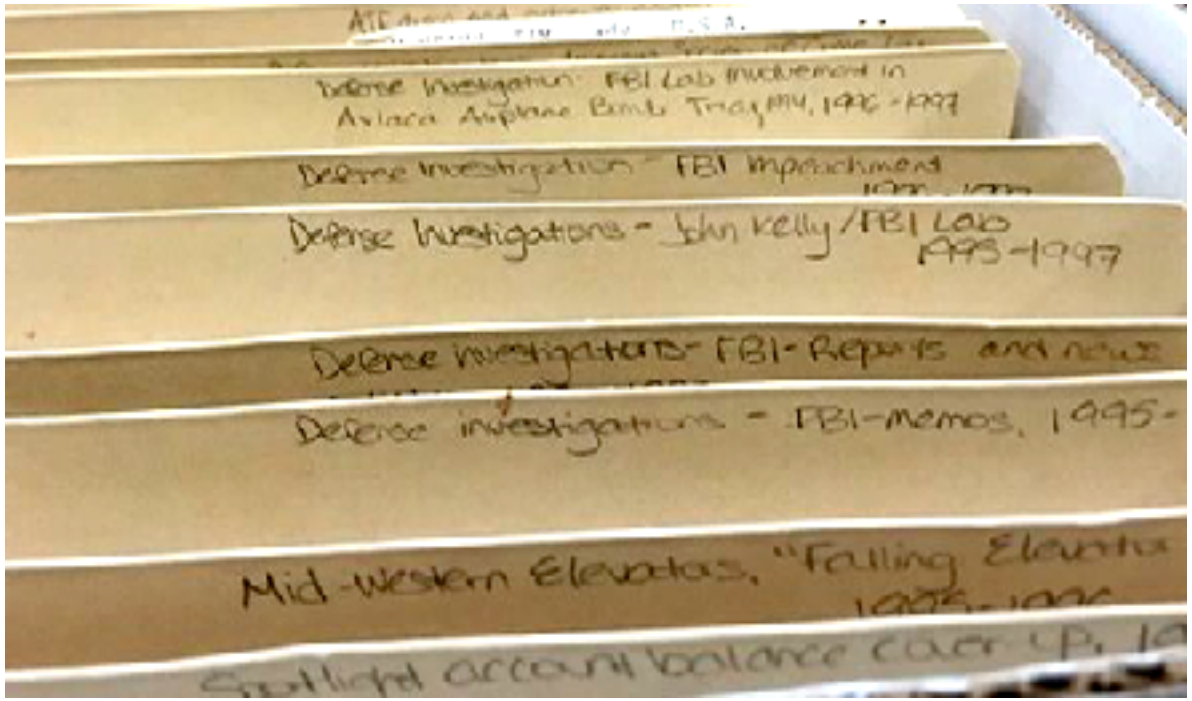
What is the best way to preserve an email collection?



TOMES

Can 35 million records be made accessible in the traditional sense?

What is the best way to preserve an email collection?



TOMES

<NOISE />

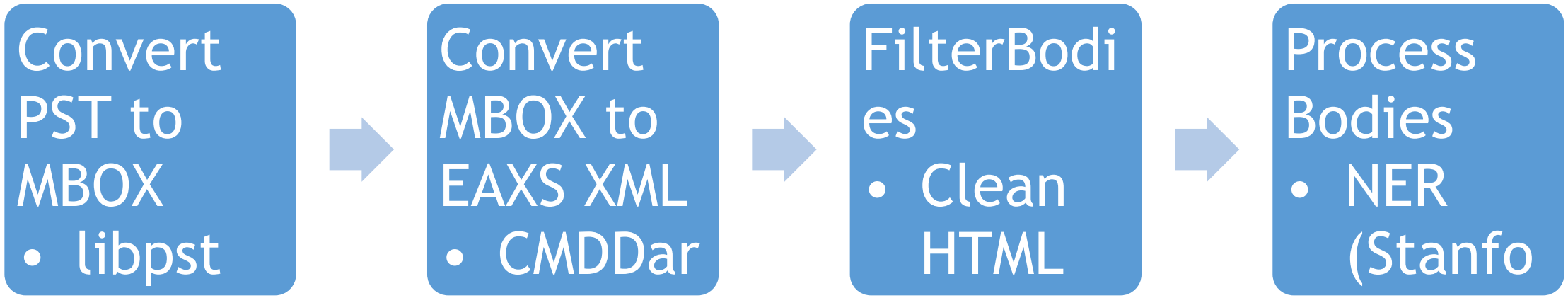

<html>
<head>
<meta http-equiv="Content-Type" content="text/html; charset=utf-8">
<style type="text/css">body,.backgroundTable{background-color:#FFFFCC;}#contentTable{border:2px solid #c0c0c0;margin-top:10px;}.headerTop{background-color:#fffaed;border-top:0px none #000000;border-bottom:1px solid #d3d3d3;text-align:center;padding:3px;}.adminText{font-size:10px;color:#999999;line-height:200%;font-family:verdana;text-decoration:none;}.headerBar{background-color:#FFFFFF;border-top:3px none #333333;border-bottom:4px none #d3d3d3;padding:1px;}.headerBarText{color:#333333;font-size:30px;font-family:Verdana;font-weight:normal;text-align:left;}.title{font-size:24px;font-weight:bold;color:#000000;font-family:Georgia,Times;line-height:150%;}.subTitle{font-size:18px;font-weight:bold;color:#000000;font-style:normal;font-family:Georgia;}.defaultText{font-size:12px;color:#333333;line-height:150%;font-family:Verdana;background-color:#FFFFFF;padding:20px;border:0px none #FFFFFF;}.footerRow{background-color:#fffaed;border-top:10px none #FFFFFF;padding:30px;}.footerText{font-size:10px;color:#333333;line-height:100%;font-family:Verdana;}a,a:link,a:visited{color:#000000;text-decoration:underline;font-weight:normal;}.headerTop a{color:#996600;text-decoration:none;font-weight:normal;}.footerRow a{color:#666666;text-decoration:underline;font-weight:normal;}body,.backgroundTable{background-color:#FFFFFF;}</style>
</head>
<body leftmargin="0" marginwidth="0" topmargin="0" marginheight="0" offset="0" style="background-color: #FFFFFF;">

<table width="100%" cellspacing="0" class="backgroundTable" style="background-color: #FFFFFF;">
<tr>
<td valign="top" align="center">

<table id="contentTable" cellspacing="0" cellpadding="0" width="550" style="border: 2px solid #c0c0c0;margin-top: 10px;"><tr><td>
<table width="550" cellpadding="0" cellspacing="0">
<tr>
<td class="headerTop" align="center" style="background-color: #fffaed;border-top: 0px none #000000;border-bottom: 1px solid #d3d3d3;text-align: center;padding: 0px;"> Email not displaying correctly or links not working? View it in your browser.
 </td>
</tr>

<tr>
<td class="headerBar" style="background-color: #FFFFFF;border-top: 0px none #333333;border-bottom: 4px none #d3d3d3;padding: 1px;">
<a href="http://list-manage.us1.list-manage.com/track/click?
u=0f1ae3af4ac154c9d11cd534b&id=9bbd860b82&e=914e7e6737" style="color: #666666;text-decoration: underline;font-weight: normal;"><img src="http://gallery.mailchimp.com/0f1ae3af4ac154c9d11cd534b/images/sgg6_4c_planned_Copy_4.jpg" alt="Southern Growth" border="0" style="margin: 0; padding: 0;">
</td>
</tr>

TOMES


Convert PST to MBOX
• libpst
Convert MBOX to EAXS XML
• CMDDar
FilterBodies
• Clean HTML
Process Bodies
• NER (Stanfo

EAXS

```
...
<SingleBody>
    <ContentType>text/plain</ContentType>
    <Charset>windows-1252</Charset>
    <BodyContent>
        <Content><![CDATA[Yes, I'll be attending nlp4arc.
        Where are we going for lunch? :-)
        - J

        ---
        John Doe
        State Archives of North Carolina]]>
</Content>
    </BodyContent>
</SingleBody>
...
```

Tagged EAXS

```
<SingleBody>
    <ContentType>text/plain</ContentType>
    <Charset>windows 1252</Charset>
    <BodyContent>
    <Content>
    Attached you can see the statistics out of the Democratic news paper <Org>The News & Observer</Org>.
    It is looking for change with the ABC set up and showing you that health care is not a high priority.
    How long did it take the legislature with a questionable vote to get the lottery into <Loc>North Carolina</Loc> even
    with polls showing a 70% of the people wanted it. How many millions of dollars in tax money did we loose from
    all of the states around us before fruit jar Baptist came out of the closet.  I would hope that the ABC
    system can be done away with in a quick fashion.  The <Loc>State of North Carolina</Loc> does not need to be in
    the retail business. If the ABC system is so great let it take over beer and wine also.  Let's get out of the dark
    ages and quit funding people like <Per>Billy Williams</Per> with tainted money.

    Election year is already on us and health care is already in place for everyone.  You and I are paying for it right now.
    Go to the emergency room at <Org>Wake Med</Org> and you can see it alive and well.  In my business I offer my
    employees <Org>Blue Cross & Blue shield</Org> for $20 per week and I still have employees turn it down.

    <Per>Obama</Per> is not the answer to every problem.

      Regards,

      <Per>John Holland</Per>

    </Content>
    </BodyContent>
</SingleBody>
```

References

EAXS Schema	http://www.history.ncdcr.gov/SHRAB/ar/emailpreservation/mail-account/mail-account_docs.html
Stanford NLP	http://nlp.stanford.edu/
spaCy NLP	https://spacy.io/
GATE NLP	https://gate.ac.uk/
NLTK	http://www.nltk.org/
TOMES GitHub	https://github.com/StateArchivesOfNorthCarolina/TOMES