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
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 ...
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...
Approaches to Email

Stanford University Libraries
- ePADD
NLP to the rescue ...
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
Can 35 million records be made accessible in the traditional sense?

What is the best way to preserve an email collection?
Can 35 million records be made accessible in the traditional sense?

What is the best way to preserve an email collection?
Convert PST to MBOX
• libpst
Convert MBOX to EAXS XML
• CMDDar
Filter Bodies
• Clean HTML
Process Bodies
• NER (Stanfo
...<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>...
Attached you can see the statistics out of the Democratic newspaper The News & Observer. It is looking for change with the ABC act 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 North Carolina even with polls showing a 70% or the people wanted it. How many millions of dollars in tax money did we lose because of all 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 State of North Carolina 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 Billy Williams with taunted 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 Wake Med and you can see it alive and well. In my business I offer my employees Blue Cross & Blue Shield for $20 per week and I still have employees turn it down.

Regards,

John Holland
## References

<table>
<thead>
<tr>
<th>Reference</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>EAXS Schema</td>
<td><a href="http://www.history.ncdcr.gov/SHRAB/ar/emailpreservation/mail-account/mail-account_docs.html">http://www.history.ncdcr.gov/SHRAB/ar/emailpreservation/mail-account/mail-account_docs.html</a></td>
</tr>
<tr>
<td>Stanford NLP</td>
<td><a href="http://nlp.stanford.edu/">http://nlp.stanford.edu/</a></td>
</tr>
<tr>
<td>spaCy NLP</td>
<td><a href="https://spacy.io/">https://spacy.io/</a></td>
</tr>
<tr>
<td>GATE NLP</td>
<td><a href="https://gate.ac.uk/">https://gate.ac.uk/</a></td>
</tr>
<tr>
<td>NLTK</td>
<td><a href="http://www.nltk.org/">http://www.nltk.org/</a></td>
</tr>
<tr>
<td>TOMES GitHub</td>
<td><a href="https://github.com/StateArchivesOfNorthCarolina/TOMES">https://github.com/StateArchivesOfNorthCarolina/TOMES</a></td>
</tr>
</tbody>
</table>