

Lessons Learned: Archiving E-Mail

Collaborative Electronic Records Project (CERP)

Midwest Archives Conference, 2008

THE ROCKEFELLER ARCHIVE CENTER



Smithsonian Institution Archives

The Basic Lessons

- Archiving for history is unlike archiving for Sarbanes-Oxley
- Email standards aren't (and may never be)
- Volume and scale is essentially unlimited
- Native email formats aren't forever
- DSpace and other open source archival tools need to be optimized for the peculiarities of email
- Yet our working prototype shows that these issues can be surmounted

This documentation is released by the Collaborative Electronic Records Project under a Creative Commons Attribution-Noncommercial-Share Alike 3.0 United States License, 2008 and 2009. This license can be viewed at <http://creativecommons.org/licenses/by-nc-sa/3.0/us/>

Citation

Burbeck, Steve Dr. (2008). "Lessons Learned: Archiving E-Mail." Sleepy Hollow, NY and Washington DC: The Collaborative Electronic Records Project.

Why Not Commercial Solutions?

- Historical archives aim at very long term preservation. Commercial solutions aim at the earliest possible legal destruction of email.
- Historical archives cannot depend for decades upon **any** proprietary software supplier, operating system or application
- For the long term, email message bodies must be converted to and stored in an open, self-describing format
- Note: email attachments present related preservation problems that have been addressed by other digital document archiving projects, e.g., the Harvard JHOVE project.

The Storage Format - XML

- Why not just use Native email format?
 - Which one? How well is it documented? How long will software exist to read it? Which companies (if any) have a real commitment to stability and longevity?
- Why eXtensible Markup Language (XML)?
 - XML is open, human readable and “self describing”
 - A good descriptive schema allows validity checking
 - There are many open source tools to create, manipulate and read XML

The Importance of a Common Schema

- A Schema defines how the tags that describe the many various parts of an email relate to each other.
 - <Account>, <Folder>, <Message>, <Header>, <Body>, <Attachment>, etc.
- The 'Mail-Account' XML schema which serves the purposes of both CERP and EMCAP (thanks to David Minor of the NC State Archives)
- It's the Rosetta stone that guides how raw email is prepared and converted to XML
- ...and it defines the start point for subsequent search, display, provenance, preservation, etc.
- It will be made public, so you don't have to reinvent the wheel

Don't Email "Standards" Make it Easy?

- The simple answer: NO
- Email evolved for several years before the first standards were developed.
- Evolution of email continues and standards continue to lag.
- Standards usually must support virtually all preexisting practices...a nearly impossible goal.
- Resulting standards tend to be "loose" and can often be interpreted in multiple (and surprising) ways

Variety is the Spice of Email

- The dozens of common email systems are not completely interoperable
 - We have tested mail from at least two dozen clients including Outlook/Exchange, Thunderbird, AOL, Eudora and AppleMail. Each has its peculiarities.
- Some use non-standard date formats
- Non-ASCII (actually, non UTF-8) characters in European and Asian mail
- Problematic HTML – older email may have HTML in inappropriate places
- Forwarded and other “child” messages may be included in nonstandard forms

Other Challenges

- Security – archives should attempt to detect and neutralize viruses and other malware, and separate out spam when possible
- Scale – one person’s inbox may have tens of thousands of messages and gigabytes of storage. A challenge for the tools
 - For example, validating a gigabyte XML file crashes some XML tools and can be very slow even if the tool doesn’t crash.

Prototype Email Conversion Results

- We have converted and validated 70 thousand messages in three test sets to the XML Mail-Account schema
 - Smithsonian - 5,537 messages in 232 Mb of recent Outlook mail
 - 99.97% successfully parsed (4 unparsed),
 - Smithsonian - 28,000 messages in 1.5 Gb Outlook account
 - 99.975% successfully parsed (5 unparsed)
 - Rockefeller Archives - 43,778 messages in 378 Mb of older eclectic mail for RAC
 - 99.85% successfully parsed (74 unparsed, but improvement is clearly possible)

Lessons Learned

- 100% success is an unrealistic goal
- We **can** achieve at least 99.9% success (and save the few unparsed emails for human inspection)
- And DSpace can store and retrieve it