Preserving Email: The Nature of the Problem

Christopher J. Prom, Ph.D
Assistant University Archivist and
Associate Professor of Library Administration
prom@illinois.edu

Digital Preservation Coalition Briefing
Wellcome Collection Conference Center, London
July 29, 2011
Googling. .

[Google search results for "email preservation guideline"]

- Email Preservation Guide Teaser | Practical E-Records
  - e-records.chrisprom.com/?p=2013 - Cached
  - Email Preservation Guide Teaser. On July 6, 2011, in Research, by Chris Prom. Here's the first draft of the intro to my upcoming report on email...

- Email Preservation and the Law of Unintended Consequences...
  - e-records.chrisprom.com/?p=2028 - Cached
  - Jul 12, 2011 – Yesterday, I sent the first rough draft of my guide to...

- Email Management and Preservation Guidelines | Practical E-Records
  - e-records.chrisprom.com/?page_id=1301 - Cached
  - The information presented below is intended to be used as a template for an...

[Additional search results and options]
A Twelve Step Plan?
Step One
Step Three . . .

- \(;AS^T\downarrow\downarrow\rightarrow\rightarrow\rightarrow\rightarrow\rightarrow S[\text{Enter}]\)
Step Twelve?

• Having had a spiritual awakening as the result of these steps, we tried to carry this message to email-holics, and to practice these principles in all our affairs.
Reason One: What Email Is

• As technology it is a:
  – Saturated
  – Interwoven
  – Commonplace
  – Malleable
  – Embedded . . .

• Utility, which

• Leaves behind evidence. . .
Email as Evidence

-Man vs. Man
Reason two: Tech

- Communicated information = A record
- Interaction of Mail Transfer Agents and User Agents
- Flexible/extendable headers, body, and content
- MIME = Multipurpose Internet Mail Extensions
- Embedded formats and references
- What are the significant properties?
  - [http://www.significantproperties.org.uk/email-testingreport.html](http://www.significantproperties.org.uk/email-testingreport.html)
- No standard storage format for msgs or MIME
  - Many binary formats, styles, etc.
  - Where’s Wally?
(Tech positives)

• Transmission standardization
• Move to server based storage and IMAP
• MBOX as quasi standard
• Ability to develop storage standard.
Reason three: Legal context

• Incentives to keep email
• Incentives to destroy email
• Discovery rules—the wildcard, nation specific
Reason four: Institutional Factors

- High cost
- Low (perceived) benefit to keep
- Risk management outlook
- How to winnow?
- Why bother?
  - Quoting an academic . . .
- Result: It’s all (usually) on the end user
The present (and future?) of email preservation
Policy: Does it work?

• Typically addresses:
  – Ownership, access rights, privacy
  – Quotas, storage, personal usage
  – Saving (where to), use of other accounts
  – Reference to other policies

• Minimal guidance

• Bottom line: It does not work to change behavior, **may** help us design better systems
Three current technical approaches

• Sweep up the crumbs
  – Guide the user
  – migrate at . . . when exactly??

• Tag it and bag it
  – ERM-driven approach

• Capture carbon . . .
  – and hope we can mine it)
Sweeping it up: some brooms

- Mailstore home
- Read pst (command line tool)
- Xena
- Follow up: InSPECT report recommendations
A Vacuum
A few XML ‘dustpans’

• Java Aperture Library (XML RDF)
• Antwerp City Archives format
• Australian National Archives (XENA)
• PeDALS email extractor
XML Account Schema

- [http://www.records.ncdcr.gov/emailpreservation/mail-account/mail-account_docs.html](http://www.records.ncdcr.gov/emailpreservation/mail-account/mail-account_docs.html)
- Stores all email for single account
- Could be used as storage system for user agent
- Multiple options for handling unicode (embed or convert)
- Extensive text and MIME handline possibilities (leave as original, convert to binhex, save externally, etc)
- Extensible headers
  - `<name> <value>` pairs
- Could write custom format via Aid4Mail scripting
- **Email Migration Tools**

  E-Mail Account Parsing

  All Account directories to be parsed -- and no subdirectories other than properly formed Account directories -- should be located in a directory named "Email Accounts" which itself must be located in the same directory as the Parser software. Each Account directory must contain all folders/subdirectories that you wish included in the parse. Examples might include Inbox and/or Sent folder. Any folder may itself contain subdirectories representing sub-folders. Within any folder or subfolder, the file containing email messages to be parsed (an .mbox file per folder subdirectory) must be in "mbox" format, with extension "mbox".

  Choose the account directory you wish to parse from the following drop-down list of available candidate accounts that appear to be well-formed.

  Once you have chosen the desired target account, press the "Proceed with parsing" button. If the chosen account has already been parsed, you will be asked whether or not you wish to reparse it.

  **Choose Account:** NO ACCOUNTS

  **Current Parse Status:** Refresh Status

  No parse status available.
Email Account Schema Overview

```xml
<?xml version="1.0" encoding="UTF-8"?>
<Account xmlns="http://www.archives.ncdcr.gov/mail-account"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.archives.ncdcr.gov/mail-account.xsd">
  <GlobalId>707093423.Account.fake.CERPHandleServer@CERP.org</GlobalId>
  <Folder>
    <Name>Hornaday_William</Name>
    <Folder>
      <Name>Inbox</Name>
      <Folder>
        <Name>BarnumPT</Name>
      </Folder>
      <Folder>
        <Name>Bison Project</Name>
        <Message>the first message</Message>
        <Message>the second message</Message>
        <Mbox>the submitted MBOX file</Mbox>
      </Folder>
    </Folder>
    <Folder>
      <Name>Expeditions</Name>
      <Folder>
        <Name>Ceylon</Name>
      </Folder>
    </Folder>
  </Folder>
</Account>
```
Classify It

• Alfresco White Paper: Total Cost of Ownership for Enterprise Content Management

• A corporate archivist’s perspective

• MeMail Project:
Carbon Capture

• Auto blindcc
• Email archiving software market
• What it does
  – Single instance storage
• Unknowns:
  – Cost (Forrester report)
  – format
  – ability to permanently preserve
  – access outside of existing infrastructure
The Access Elephant

• Copyright/ Third Party IP
• Search, Discovery, Retrieval
• Fedora and other repositories
  – Hydra Project. Need
    • content models
    • Deep search (Lucene Solr or similar)
    • Front end (Blacklight)
Sarah’s inbox: an access model?
Two Fundamental Challenges

• Building a research and development agenda:
  – User behavior, policy, standards (build on InSPECT significant properties report)

• Building tools to acquire, preserve, and make email useful for long-term (cyber-infrastructure)
  – Capture, storage, conversion, metadata, access

• Making the case to funders and potential donors
Personal ‘Archiving’

• Cathy Marshall “Rethinking Personal Digital Archiving”

• http://www.thedigitalbeyond.com/

• Lifestream concept (Eric Freeman and David Gelernter)

• Services:
  – Carbonite, Crashplan, Mozy, etc.
  – Backupify, Think Up (Gina Trapani)
A Modest Proposal

• Provide the users (and institutions) something of value *given their ‘piling’ behaviors*
  – Backup Services, *plus*
  – Think-up like services, *plus*
  – Trust, *plus*
  – *the ability to donate!*
  – [http://www.iKive.com](http://www.iKive.com)

• Investing users and funders in the problem?
Questions and Discussion
Preserving Email: The Nature of the Problem

Christopher J. Prom, Ph.D
Assistant University Archivist and Associate Professor of Library Administration
prom@illinois.edu

Digital Preservation Coalition Briefing
Wellcome Collection Conference Center, London
July 29, 2011