Preserving Email: Why to Do So (and How)

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

CARLI Webinar
May 1, 2012

Why Preserve: What Email Is

• As technology it is a:
  – Saturated
  – Interwoven
  – Commonplace
  – Malleable
  – Embedded . . .
• Utility, which
• Leaves behind evidence. . .
Email as Evidence

- Man

Googling...
A Twelve Step Plan?

Step One
Step Three . . .

- ;AS^T↓↓→→>S[Enter]

Step Twelve?

- “Having had a spiritual awakening as the result of these steps, we tried to carry this message to email-aholics, and to practice these principles in all our affairs.”
Why ‘hard’ to preserve: 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 messages or MIME content (attachments)
  - Many binary formats, styles, etc. (mbox, eml, pst, proprietary/closed formats)
(Tech positives)

- Transmission standardization
- Increased use of server based storage, IMAP
- MBOX as quasi standard
- Ability to develop a storage standard.

Why hard to preserve: legal context

- Incentives to keep email
- Incentives to destroy email
- Discovery rules—the wildcard, nation specific
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
Option 1: 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 better options

- Bag it and tag it
  - ERM-driven approach
- Sweep up the crumbs/nurture and harvest
  - Capture at end of life, or
  - Guide the user and migrate at future point
- Capture carbon . . .
  - and mine it later
Tools: Bag It and Tag It

- Alfresco White Paper: Total Cost of Ownership for Enterprise Content Management
- A corporate archivist’s perspective
- Simpler version: MeMail Project
  - University of Michigan

Sweeping it up: some brooms

- See InSPECT significant properties report
  - http://www.significantproperties.org.uk/testingreports.html
- Tools:
  - Mailstore Home
  - Xena
  - Read pst (command line tool)
  - Emailchemy
  - Project Muse
  - Aid4Mail
MailStore Home

- Free
- Windows Client
- Downloads to Local Computer
- Must be run manually
- Dependent on scheduling and backup
- Personal Example of server-based tools

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 handling possibilities (leave as original, convert to binhex, save externally, etc)
- Extensible headers
  - `<name> <value>` pairs
- Could write custom format via Aid4Mail scripting
CERP (Collab. Electronic Records Project) Parser

CERP: Email Account Schema Overview

```xml
<folder>
    <name>Homayouz_Hamids</name>
    <folder>
        <name>lnbox</name>
        <folder>
            <name>Mailbox</name>
            <folder>
                <name>Isman Project</name>
                <message>
                    <subject>Message 1</subject>
                    <body>First message body</body>
                </message>
                <message>
                    <subject>Message 2</subject>
                    <body>Second message body</body>
                </message>
            </folder>
        </folder>
    </folder>
</folder>
```
Case Studies

- Harvard
  - [www.ifstuwien.ac.at/dp/ipres2010/papers/goethals-08.pdf](http://www.ifstuwien.ac.at/dp/ipres2010/papers/goethals-08.pdf)
- Oxford/Bodleian

Project Muse

- Sudheendra Hangal
Tools: 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

EmailArchiva

- Capture at point of transmission
- Wide server support
- Filesystem storage in .eml (RFC 2822) format
- Single-instance storage/compression
- Integrated web access and discovery system
- Retention/compliance/discovery Features
- Pending Case Study
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?
Project Muse: Visualization

Three Long-term 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.dlib.org/dlib/march08/marshall/03marshall-pt1.html](http://www.dlib.org/dlib/march08/marshall/03marshall-pt1.html).
- [http://www.thedigitalbeyond.com/](http://www.thedigitalbeyond.com/)
- Lifestream concept (Eric Freeman and David Gelernter)
- Services:
  - Carbonite, Crashplan, Mozy, etc.
  - Backupify, Think Up (Gina Trapani)

Emergent Work (In Illinois)

- 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