




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. . .

I

Email as Evidence



v.



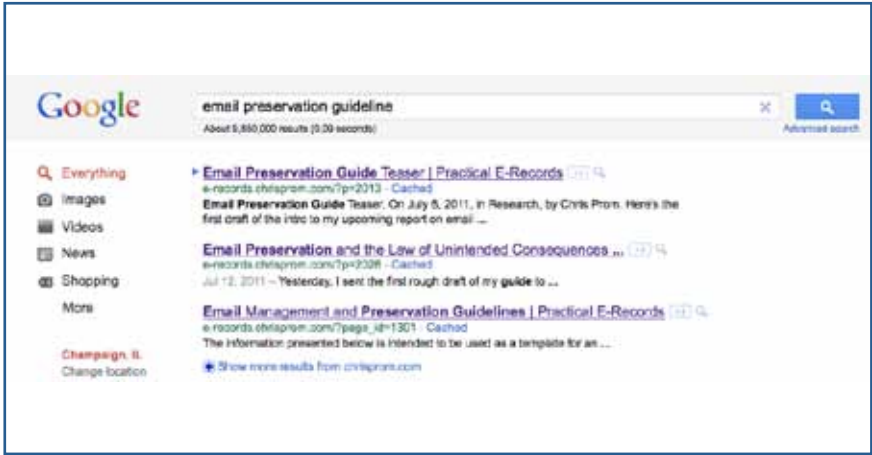
-Man

UNIVERSITY LIBRARY
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

4

I

Googling. .

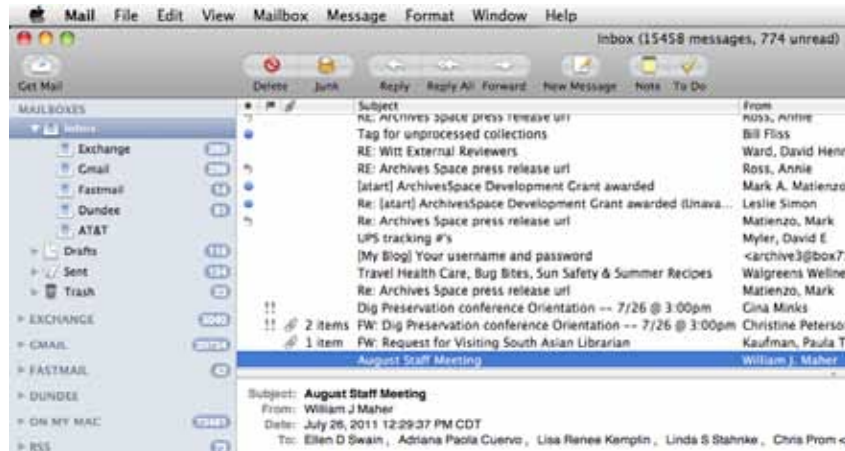


UNIVERSITY LIBRARY
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

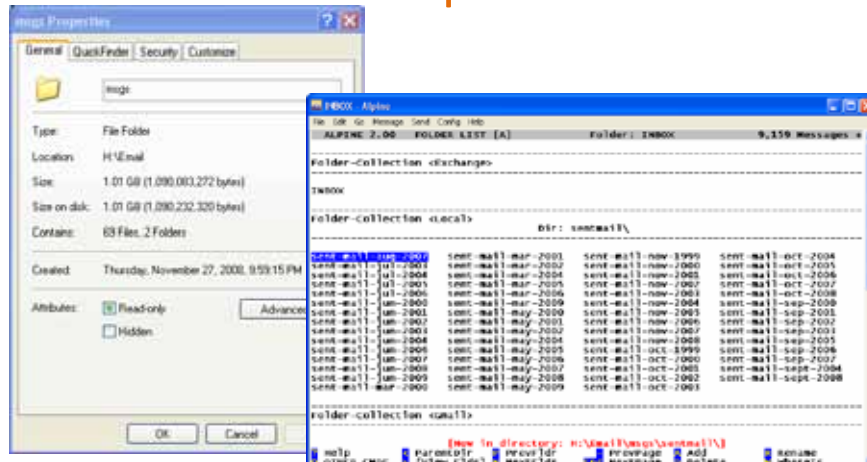
5



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.”




UNIVERSITY LIBRARY
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

10



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>
- No standard storage format for messages or MIME content (attachments)
 - Many binary formats, styles, etc. (mbox, eml, pst, proprietary/closed formats)

UNIVERSITY LIBRARY
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

11



(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

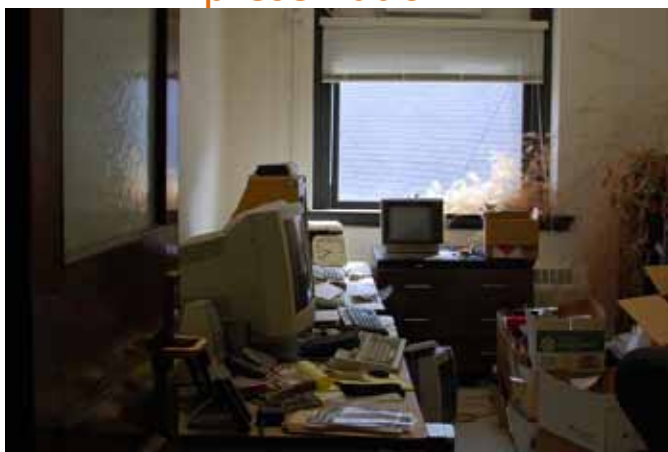


Institutional Factors

- 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
 - <http://blogs.alfresco.com/wp/democast/category/email-archive/>
- A corporate archivist's perspective
- Simpler version: MeMail Project
 - University of Michigan
 - <http://e-records.chrisprom.com/?p=1965>



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

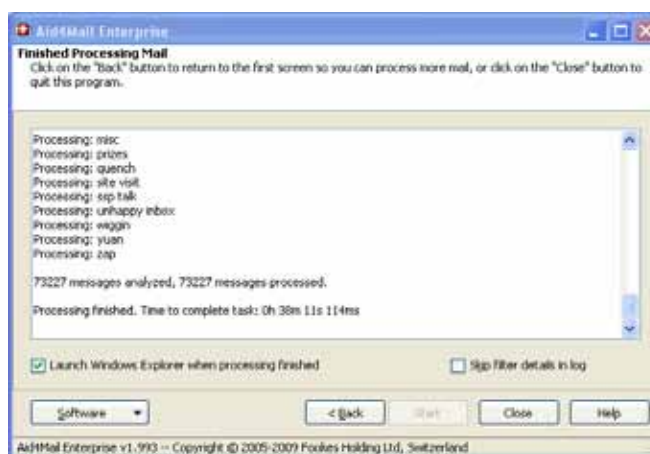


MailStore Home

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

```



Case Studies

- Harvard
 - www.ifs.tuwien.ac.at/dp/ipres2010/papers/goethals-08.pdf
- Oxford/Bodleian
 - <http://e-records.chrisprom.com/?p=2200>



Project Muse

- <http://mobisocial.stanford.edu/muse/>
- Sudheendra Hangal
- archive.org/details/personaldigitalarchiving2012pt1



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
 - <http://e-records.chrisprom.com/?p=2215>



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?

The screenshot shows an email inbox for Sarah. The interface includes a search bar, a list of folders on the left, and a main list of email messages. Each message entry shows the sender, subject, and date. Some messages are highlighted in blue, indicating they are selected or unread.

Sender	Subject	Date
Gov. Sarah Palin, Kelly C. Goode (GOV), Kelly C. Goode	Walter joined to resign	01/03/09
govin@inspirelib.com	PHS and SES investment funds	10/06/08
Palin, Gov. Sarah Palin, Kira Perry	Governor's Press Release	08/20/08
Gov. Sarah Palin, Palin	Governor's Press Release	08/20/08
Brad Frazeech, Gov. Sarah Palin	One Akeke Housing Project REE	08/20/08
Gov. Sarah Palin, Kira Perry, Michael A. Vanden (GOV), Palin	PHS Acute Careline Decision - CONFIDENTIAL ATTORNEY-CLIENT	08/03/08
Gov. Sarah Palin	Legal and medical	06/23/08
Beth Leachman, Sarah Palin, Gov. Sarah Palin, insarah@outlook.com	Health care for Akeke	06/23/08
Gov. Sarah Palin	From Dr. Louise Holtz www.PainRelief.com	06/19/08
Gov. Sarah Palin	SEC in my house: reasons why the Storage Boat project	06/19/08
Gov. Sarah Palin (2)	Salary gift from Waga Waga	06/19/08
Katey Palin	New email address not working	06/19/08
Gov. Sarah Palin, Kira Perry, Michael A. Vanden (GOV), HG (2)	PHS: Nurses Meet Opioid, Pain Strategy and the economy	06/19/08
Gov. Sarah Palin	A Difficult Decision	06/19/08
Beth Leachman, Sarah Palin, insarah@outlook.com	Delivery Notification Delivery has failed	06/17/08
Beth Leachman, Sarah Palin, Chuck Heath, insarah@outlook.com	Spread the word	06/17/08
Gov. Sarah Palin, Lorna Smith, Michael A. Vanden (GOV)	PHS: Can't Hear Congressman Young's Office? Part Taken full issue?	06/17/08
Gov. Sarah Palin (2)	Delivery Notification: Delivery has failed	06/17/08
Cara Clavin, Gov. Sarah Palin, Michael A. Vanden (GOV), Sharon Longman	Re: Sara Mable	06/17/08
Beth Leachman, Chuck Heath, insarah@outlook.com, Todd Palin	Stannah's Stannah...	06/17/08



Project Muse: Visualization

UNIVERSITY LIBRARY
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

32



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

UNIVERSITY LIBRARY
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

33



Personal 'Archiving'

- Cathy Marshall “Rethinking Personal Digital Archiving”
 - <http://www.dlib.org/dlib/march08/marshall/03marshall-pt1.html>.
- <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>
- Investing users and funders in the problem