

Preserving Email: The Nature of the Problem

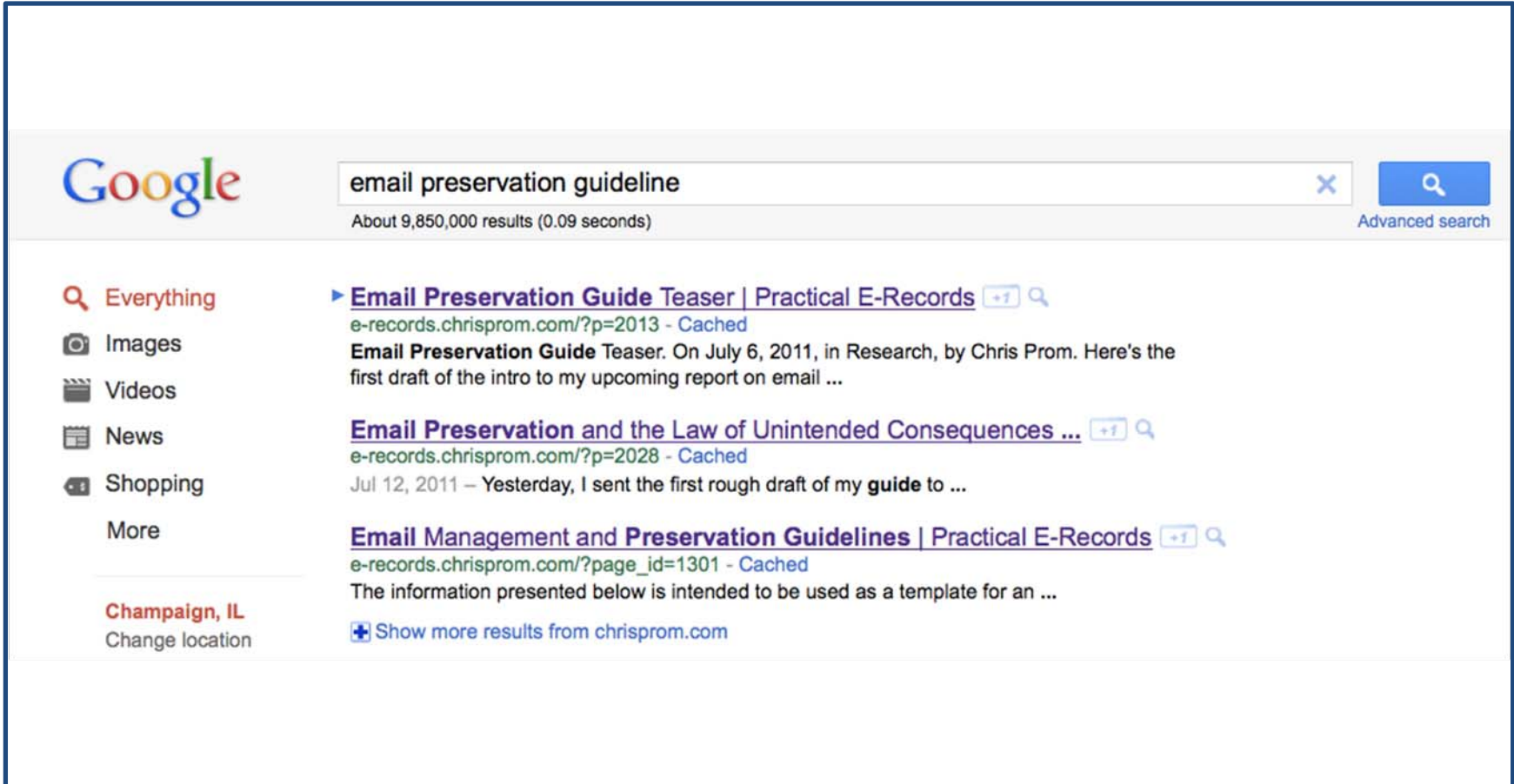
Christopher J. Prom, Ph.D
Assistant University Archivist and
Associate Professor of Library Administration
prom@illinois.edu

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



UNIVERSITY LIBRARY
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Googling. .









The screenshot shows a Google search interface. The search bar contains the text "email preservation guideline". Below the search bar, it says "About 9,850,000 results (0.09 seconds)". To the right of the search bar is a blue button with a magnifying glass icon and the text "Advanced search".

On the left side of the page, there is a vertical menu with the following items:

- Everything
- Images
- Videos
- News
- Shopping
- More

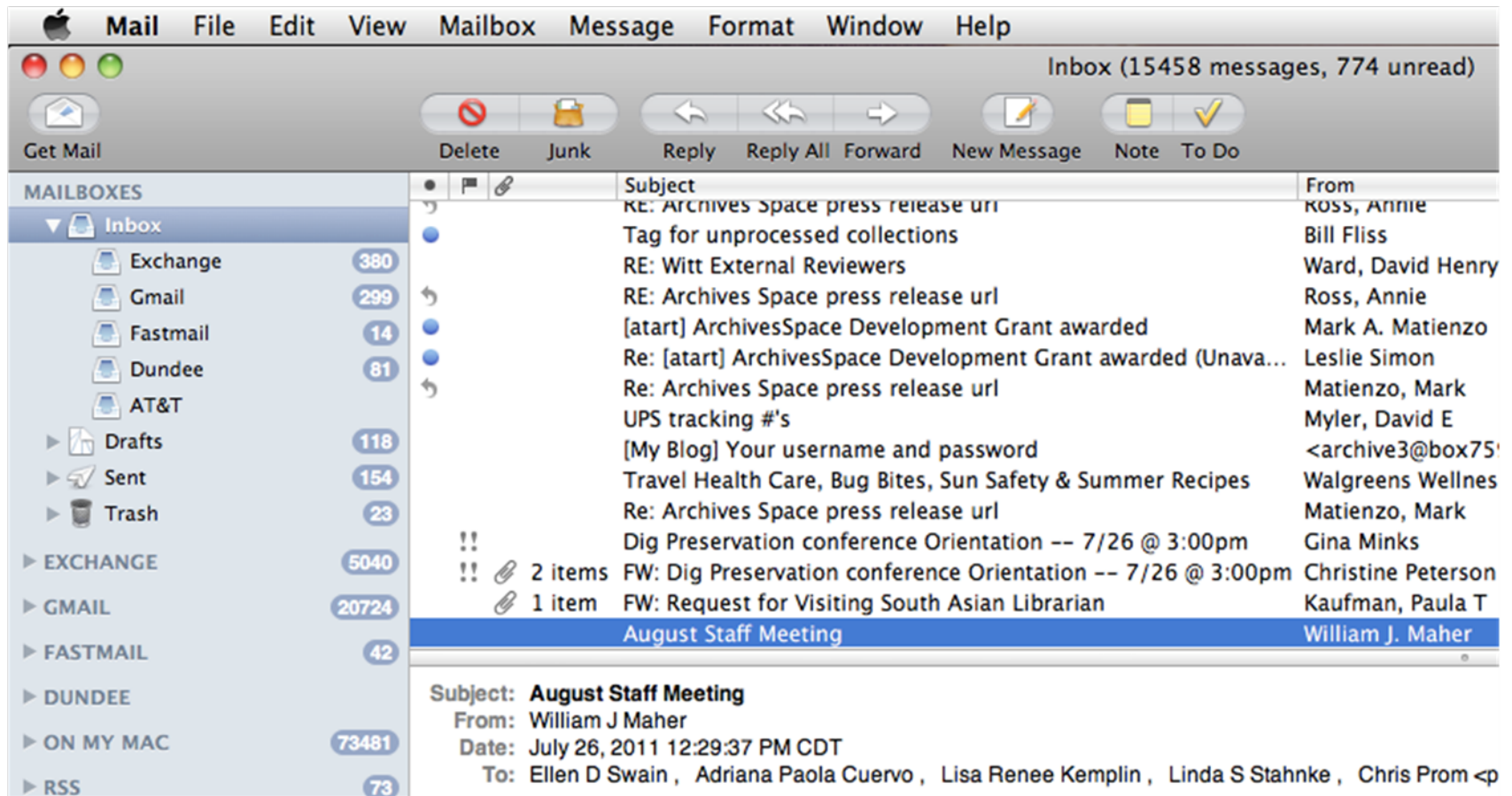
Below the menu, it says "Champaign, IL" and "Change location".

The search results are as follows:

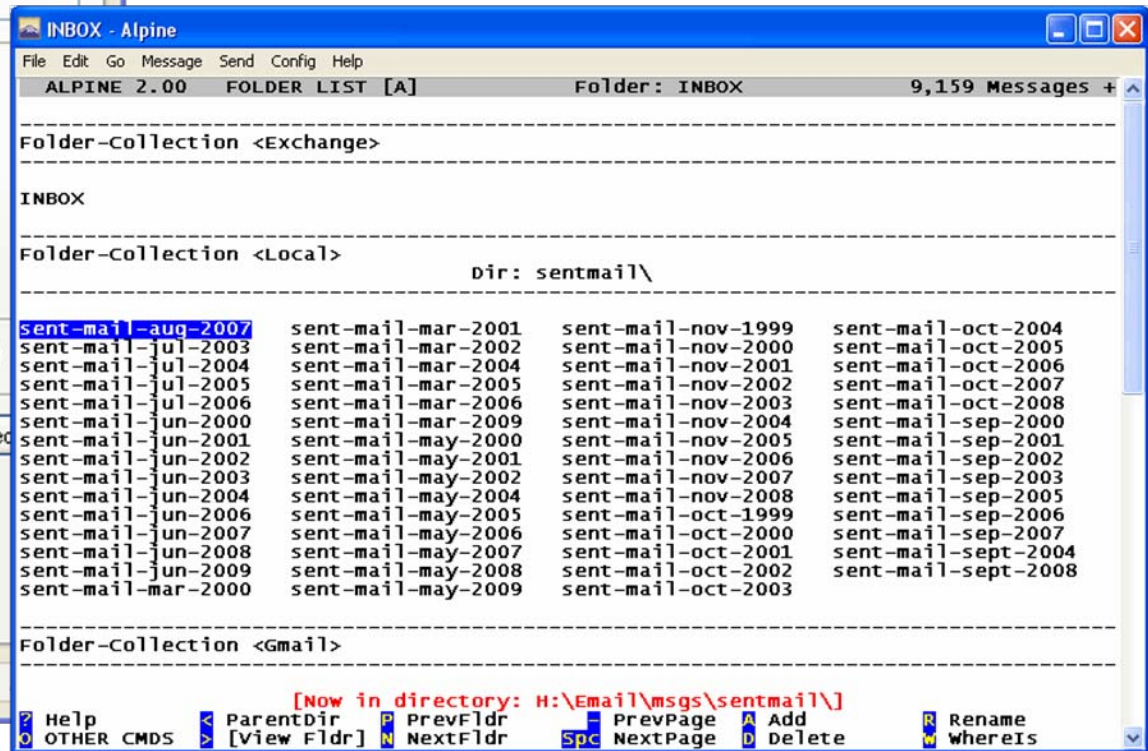
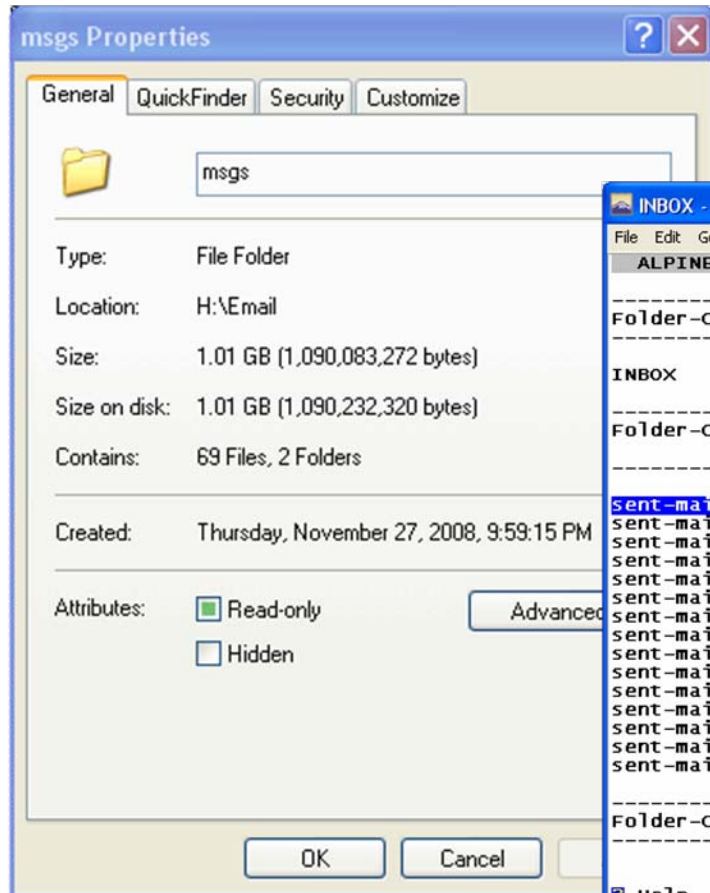
- Email Preservation Guide Teaser | Practical E-Records**  
e-records.chrisprom.com/?p=2013 - [Cached](#)
Email Preservation Guide Teaser. On July 6, 2011, in Research, by Chris Prom. Here's the first draft of the intro to my upcoming report on email ...
- Email Preservation and the Law of Unintended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

At the bottom of the results, there is a link: [+ Show more results from chrisprom.com](#)

A Twelve Step Plan?

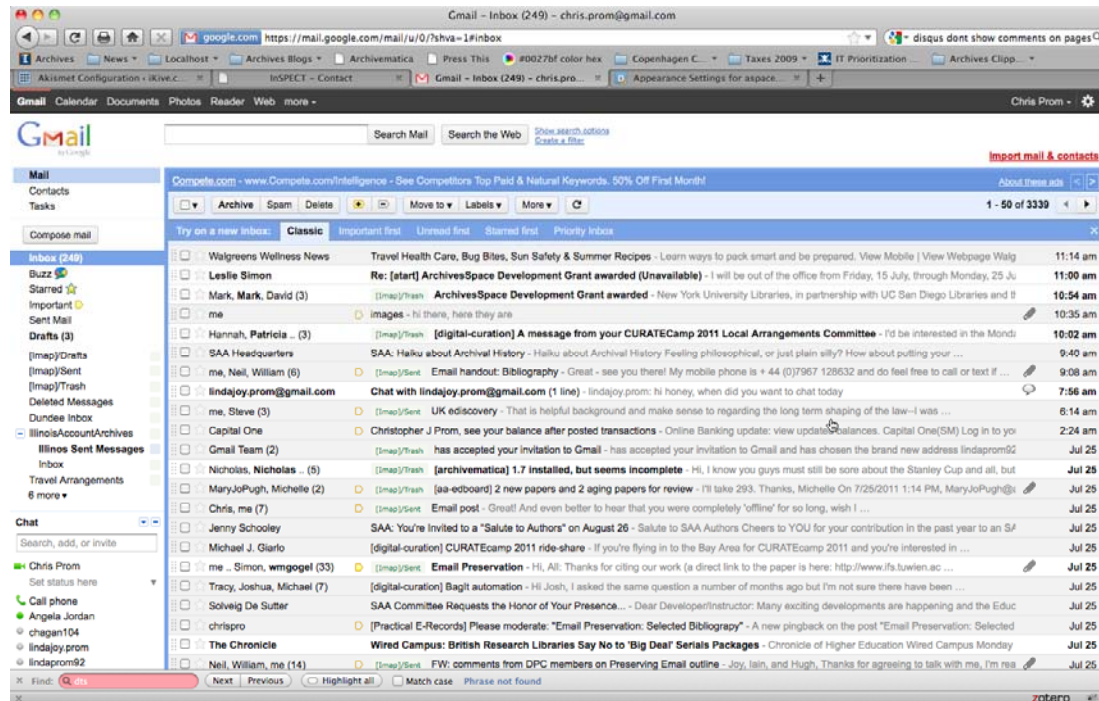


Step One



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-holics, and to practice these principles in all our affairs.



**WHERE ARE WE GOING AND WHY
ARE WE IN THIS HANDBASKET?**

Reason One: What Email Is

- As technology it is a:
 - Saturated
 - Interwoven
 - Commonplace
 - Malleable
 - Embedded . . .
- Utility, which
- Leaves behind evidence. . .

Email as Evidence



v.



-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>
- 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

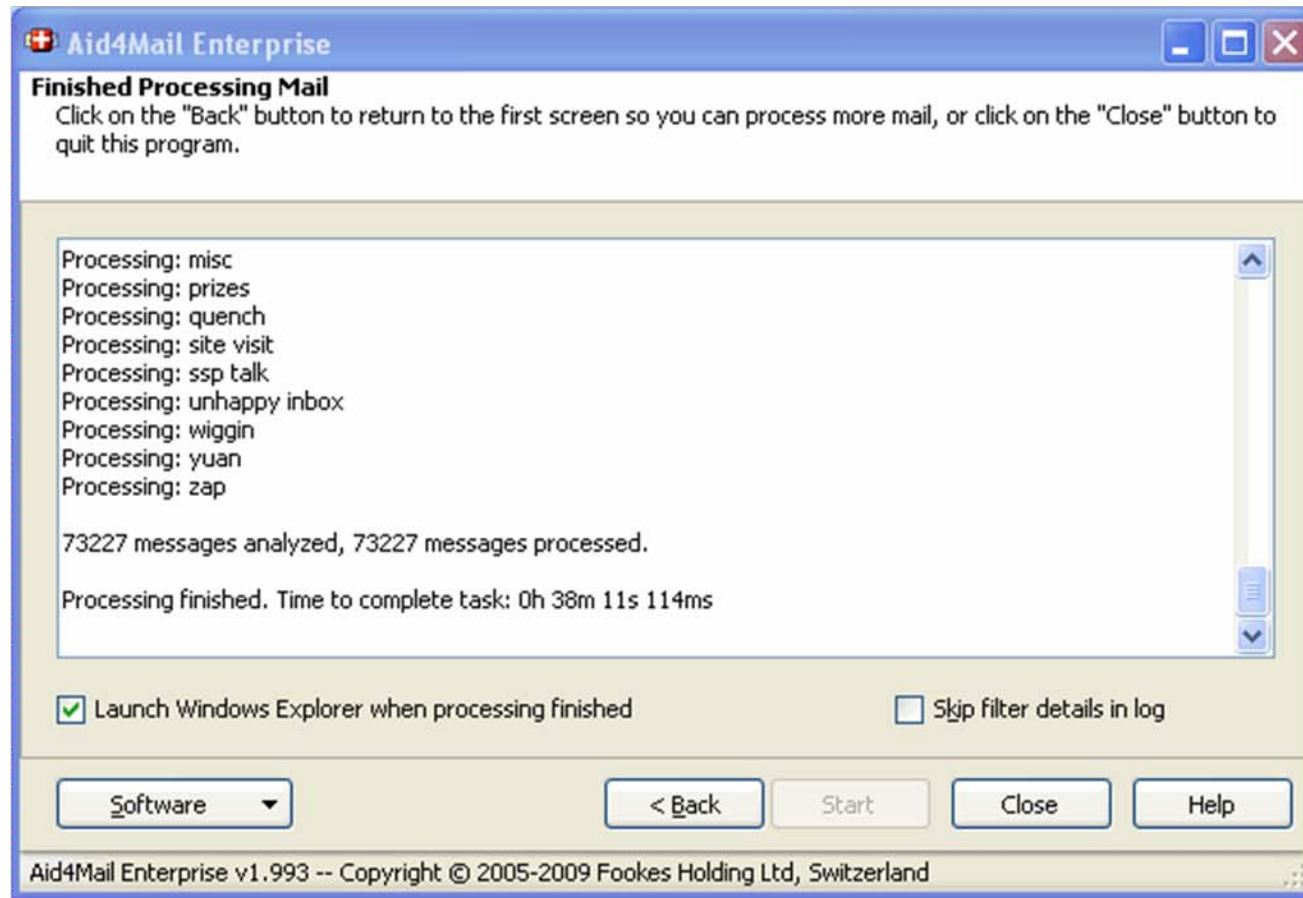
Three current technical approaches

- 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 Apeture 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

• Email

Seaside - Microsoft Internet Explorer provided by SINET

http://localhost:9091/seaside/EmailParsing

File Edit View Favorites Tools Help Links EmailParsing SINetFromHome webTA SnagIt

Seaside

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 folder subdirectories that you wish included in the parse. Examples might include InBox and/or Sent folders. Any folder may itself contain subdirectories representing sub-folders. Within any folder or subfolder, the file containing email messages to be parsed (one 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.

[See help pages for more detail](#)

Choose Account:

Current Parse Status

No parse status available

New Session Configure Toggle Halos Profiler Memory Terminate XHTML 12/0 ms

Done Local intranet 100%

Email Account Schema Overview

```
<?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>
        <Name>Expeditions</Name>
        <Folder>
          <Name>Ceylon</Name>
        </Folder>
      </Folder>
    </Folder>
  </Folder>
</Account>
```

Classify 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
- MeMail Project:
 - <http://e-records.chrisprom.com/?p=1965>

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?

A project of **Sunlight Foundation** (Who is the Sunlight Foundation?) [Join Our Mailing List](#)

[Home](#) Sarah Palin | [Sign out](#)



Sarah's Inbox
A project of the Sunlight Foundation

Inbox

Sent Mail

Drafts

All Mail

Starred (all)

Starred (by you)

Sample Searches

flippinbelieveit

"bridge to nowhere"

"Tina Fey"

"horrible people are bringing you down"

"Alaska natural gas pipeline"

"barack obama"

crud

"first dude"

"who leaked it?"

"What a dumbass he is"

"who's going to trim my hair?"

pizza

About

This site parses the emails sent and received by Sarah Palin while she was governor of Alaska and presents them in a more familiar interface.

To create Sarah's Inbox, we used the digitized emails released by

★ Gov. Sarah Palin, Kelly C Goode (GOV), KelyC (GOI0 tpalin@mccain08hq.com	Keller asked to resign	07/01/2009
★ 'Palin, Gov. Sarah Palin, Kris Perry	PERS and SBS investment funds	10/06/2008
★ Gov. Sarah Palin, Palin	Governor's Press Release	09/30/2008
★ Brad Fluetsch, Gov. Sarah Palin	Governor's Press Release	09/30/2008
★ Gov. Sarah Palin, Kris Perry , Michael A Nizich (GQV), Palin	Give Alaska Housing Finance \$\$\$\$	09/26/2008
★ Gov. Sarah Palin	FW: Ninth Circuit Chistochina decision - CONFIDENTIAL ATTORNEY-CLIENT PRIVILEGED	09/23/2008
★ Beth Leschper, Bristol Palin, Gov. Sarah Palin, mbkrdk@starband . net	Legal and lawsuits	09/22/2008
★ Gov. Sarah Palin	Useful info for Alaska	09/20/2008
★ Gov. Sarah Palin	from Dr. Laurie Roth www.therothshow.com	09/19/2008
★ Gov. Sarah Palin	BBQ at my house manana with the Strange Boys playing!	09/18/2008
★ Gov. Sarah Palin (2)	baby gift from Maya Wrap	09/18/2008
★ Todd Palin	New email address not working	09/18/2008
★ Gov. Sarah Palin, Kris Perry , Michael A Nizich (GQV), re: (2) Gov. Sarah Palin	FW: Tesoro Iron Dog, Flat Stanley had the scoop	09/18/2008
★ Beth Leschper, Bristol Palin, mbkrdk@starband.net	A Difficult Decision	09/17/2008
★ Beth Leschper, Bristol Palin, Chuck Heath, mbkrdk@starband.net	Delivery Notification : Delivery has failed	09/17/2008
★ Gov. Sarah Palin, Lynne Smith, Michael A Nizich (GQV)	Spread the word	09/17/2008
★ Gov. Sarah Palin (4)	shame	09/17/2008
★ Cora Crome, Gov. Sarah Palin, Michael A Nizich (GQV), Sharon Leighow	FW: Call from Congressman Young 's Office / Fort Yukon fuel issue?	09/17/2008
★ Beth Leschper, Chuck Heath, mbkrdk@starband.net, Todd Palin	Delivery Notification: Delivery has failed	09/17/2008
	Exxon Mobile	09/17/2008
	Shameful Sharon...	09/17/2008

Two Three 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'

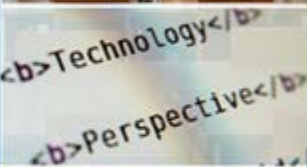
- Cathy Marshall “Rethinking Personal Digital Archiving”
 - <http://www.dlib.org/dlib/march08/marshall/03marshall-pt1.html>.
- <http://www.thedigitalbeyond.com/>
- Lifestream concept (Eric Freeman and David Gelernter)
- Services:
 - Carbonite, Crashplan, Mozy, etc.
 - Backupify, Think Up (Gina Trapani)

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>
- 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



UNIVERSITY LIBRARY
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN