Smithsonian Institution Archives’ Email Preservation Tools

What they are, how they work and their operational context

Riccardo Ferrante
Director of Digital Services
@raferrante

siarchives.si.edu
@SmithsonianArch

Lynda Schmitz Fuhrig
Electronic Records Archivist
@LyndaLSF

Archiving Email Symposium
Library of Congress
Washington, DC – June 2, 2015
Our Starting Point

Late 1980s ELM (ELectronic Mail)
Our Starting Point

Since then ... an incomplete listing
Our Starting Point

Since then ... an incomplete listing
Our Starting Point

Since then ... an incomplete listing
Our Starting Point

Since then ... an incomplete listing
Our Starting Point

Since then ... an incomplete listing
Our Starting Point

Since then ... an incomplete listing
Our Starting Point

Institutional email management strategies, circa 2004:

- SI Archives-issued guidance part of new employee orientation
- Email typically considered like voicemail
- IT deleted oldest emails on large accounts
- Email migration by special request only when switching from GroupWise to MS Outlook
- Email archives the responsibility of the person
- No established mechanism for identifying and capturing historically valuable emails

Some of the Smithsonian’s email platforms over the past 35 years.
Our Starting Point

Late 1980s – the date of some of the oldest email in the Archives collections. At the time, the Smithsonian and many universities were using ELM and PINE.

We do not refuse deposits of email on the basis of format. Therefore we must use an approach that accommodates a variety of formats.

Accessioned email comes from a variety of formats including Elm, Pine, Eudora, ccMail, Lotus Notes, Novell GroupWise and most recently Microsoft Outlook.

Some of the Smithsonian’s email platforms over the past 35 years.
The Collaborative Electronic Records Project (CERP)

With these concerns in mind, the Archives teamed with the Rockefeller Archive Center in a three-year project to explore the issue, develop guidance and workflows for acquiring and preserving email.

By the end of the project, CERP had also developed a software tool for preserving email in XML using a schema developed in collaboration with the NC/KY/PA Electronic Mail Initiative project (EMCAP).

Some of the Smithsonian’s email platforms over the past 35 years.
Objective: “Durable Access”

- Preservation using non-proprietary, standards based formats
- Asset manageability
- Retention of context
- Retention of complete messages
- Methodology with a low resource requirement
Our Approach

• Since 2008, using the CERP methodology & tools
• Focus on, and actively capturing key organization record creators
• Accounts, not individual emails
• Easier manageability at this level
CERP Model

SIP to AIP

1) Convert the collection to .mbox files, if not already in this format.
2) Use the CERP parser to convert .mbox files into an XML preservation file.
3) Assemble all components (metadata, SIP, preservation XML, and log files) into an AIP.
CERP Parser

E-Mail Account Parsing

All Account directories to be parsed — and no subdirectories other than properly named 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 outbox folders. Any folder or subfolder under or 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 extensions "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: C:EmailParser/Email_Accounts/Carl_backup_messagesave

Current Parse Status: Parsing initiated 7 January 2015 for C:EmailParser/Email_Accounts/Carl_backup_messagesave

- Begin Parsing: Calendar at 10:18:08 am
- 14 messages parsed
- Done Parsing: Calendar at 10:18:08 am
- Begin Parsing: D:AirClub Data at 10:18:08 am
- 1 messages parsed
- Done Parsing: D:AirClub Data at 10:18:08 am
- Begin Parsing: D:AirClub Data at 10:18:08 am - 28 messages parsed

New Session Configure Toggle Halos Profiler Memory Terminate XHTML 636/300 ms
Preservation and Access

Preservation output through Email Account XML Schema (EAXS)

Human and machine readable but not “pretty”
Beyond CERP

The image shows a screenshot of a software interface named DArcMail v0.1. The interface displays information about an email account named 'Carl_S'. The details include:

- **Account name**: Carl_S
- **Id**: 11
- **Account directory**: C:\Email\Parse\Email_Accts\Carl_backup_message\save
- **Message count**: 1,226
- **First message date**: 2011-05-17
- **Last message date**: 2014-05-22
- **'From' addresses**: 66
- **'To' addresses**: 97
- **'Cc' addresses**: 44
- **'Bcc' addresses**: 0
- **External content files**: 59

Below this, there is a section labeled 'Folders' with a table that lists various folders with their respective IDs, names, message counts, start dates, end dates, and relative paths.
DArcMail (Digital Archive Mail)
<table>
<thead>
<tr>
<th>Id</th>
<th>Date</th>
<th>Subject</th>
</tr>
</thead>
<tbody>
<tr>
<td>161950</td>
<td>2012-01-12</td>
<td>RE: Running Post_Init script</td>
</tr>
<tr>
<td>162490</td>
<td>2012-01-25</td>
<td>jhove2, droid, etc.</td>
</tr>
<tr>
<td>162503</td>
<td>2012-02-21</td>
<td>status</td>
</tr>
<tr>
<td>162523</td>
<td>2012-03-22</td>
<td>mysql notes</td>
</tr>
<tr>
<td>162045</td>
<td>2012-06-07</td>
<td>Jhove/Droid documentation</td>
</tr>
<tr>
<td>162595</td>
<td>2012-09-13</td>
<td>fy 11 numbers for preservation_processing folder on sia-digarch01</td>
</tr>
<tr>
<td>162610</td>
<td>2012-11-01</td>
<td>meeting notes</td>
</tr>
<tr>
<td>162174</td>
<td>2012-11-08</td>
<td>Working Guidelines for Archiving Websites and Social Media</td>
</tr>
<tr>
<td>162224</td>
<td>2013-06-13</td>
<td>FW: Transcription Center in Beta - PLEASE HELP US TEST</td>
</tr>
</tbody>
</table>
Id: 162523
Account: Car_S
Folder: Sent Items
Global Id: EB83021C560D8E4DAE5EBBC6570587590A4D862A@SI-MSEV04.US.SINET.SI.EDU
Date: Thu, 22 Mar 2012 11:10:02 -0400
Subject: mysql notes
From: "Schaefer, Carl" <SchaeferC@si.edu>
To: "Schmitz Fuhrig, Lynda" <SchmitzfuhrigL@si.edu>

Message Content

<table>
<thead>
<tr>
<th>Action</th>
<th>Storage</th>
<th>Content-Type</th>
<th>Length</th>
<th>Original Name</th>
<th>Param</th>
</tr>
</thead>
<tbody>
<tr>
<td>view</td>
<td>internal</td>
<td>text/plain</td>
<td>141</td>
<td></td>
<td></td>
</tr>
<tr>
<td>view</td>
<td>internal</td>
<td>text/html</td>
<td>1,831</td>
<td></td>
<td></td>
</tr>
<tr>
<td>download</td>
<td>internal</td>
<td>application/vnd.ms-powerpoint</td>
<td>17,783</td>
<td>mysql_install.docx</td>
<td>5</td>
</tr>
</tbody>
</table>
Here is what we did this morning. We did everything except what is in yellow.

What we installed was version 5.5.21, 64-bit.

Carl
Dear Colleagues,

The Office of Safety, Health and Environmental Management’s Occupational Health Services’ seasonal flu vaccination outreach for 2011-2012 has officially begun. On average, Occupational Health Services provides nearly 3,000 shots annually across the Smithsonian. All employees, affiliated staff, and official volunteers are strongly encouraged to get their free flu shot. The complete schedule for this year’s outreach effort can be found on the OSHEM website on PRISM:

http://ofeo.si.edu/safety_health/ohs/flu_vaccine_fy11.asp

The Centers for Disease Control and Prevention recommend a yearly flu vaccine as the first and most important step in protecting against this serious disease. While there are many different flu viruses, the flu vaccine protects against the three main flu strains that research indicates will cause the most illness during the flu season. The vaccine can protect you from getting sick from these three viruses, or it can make your illness milder if you get a different flu virus.
Encoded attachments
Lifecycle coverage from DArcMail Suite

- Processing
- Preservation
  - Everything apart from AIP assembly
- Reference Access (offline)
- Follows CERP methodology
Where this fits in the larger SIA context

**Acquisition**
Focused on capturing senior officials, notable persons on departure
Follows standard electronic records procedures
Accept unsolicited email records regardless of format
Work with OCIO to acquire the accounts

**Appraisal**
Everything received is preserved
Weeding is the responsibility of the account owner
Finalized based on subject-sender listing
*Now possible within DArcMail*

**Arrangement, Description and AIP preparation**
Subject/Sender log is a finding aid supplement
Full account structure is embedded in the preserved account’s XML.
*Now possible within DArcMail*

**Reference**
Online collection level description
“Dark archive”
Restricted for a 15-year minimum
Where CERP fits with other tools

CERP Parser handles 3 stages.

DArcMail Suite adds a UI for archival processing and works in Python.

Can be used as part of a composite suite of tools for a full lifecycle solution.

Affordable (open source) and accessible solution for preservation.