# Digital Preservation: Using the Email Account XML Schema

***Riccardo Ferrante and Lynda Schmitz Fuhrig; Smithsonian Institution Archives; Washington, DC***

## Abstract

*The Smithsonian Institution Archives (SIA) and the Rockefeller Archive Center (RAC) conducted a three-year pilot that explored preservation challenges with email collections. This paper reviews the acquisition model and workflow used based on the OAIS Reference Model. Rather than focusing on individual messages, the Collaborative Electronic Records Project (CERP) settled on preserving an account as a whole, maintaining the structure and relationships within a collection as well as simplifying metadata management. This paper also reviews some of the challenges with the email collections, including lack of organization and inclusion of non-record/sensitive material. Both archives also addressed the importance of sound recordkeeping practices and retention schedules and issued various guidance documents for depositors.*

*CERP also collaborated with another research team (the E-Mail Collaborative Initiative (EMCAP)) to develop an XML schema capable of encompassing a complete email account and its content. The E-Mail Account XML schema defines a standard XML structure for preserving an email account along with its internal organization, its messages and attachments, and the interrelationships of the messages without sacrificing granular email message data. This paper describes the schema, its unique characteristics, and its value to the archival and digital preservation communities in the context of, and comparison to, other efforts to digitally preserve email.*

*The schema structure positions preserved email accounts for multiple levels of searching strategies including: individual messages, account-wide, and cross-account search and retrieval. This helps to expose social networks and message interrelationships present in, and across, accounts.*

*The E-Mail Account schema has made possible the preservation of large bodies of related e-mail in a single XML file, as demonstrated in the recent EMCAP and CERP projects. Unlike other work in the area of e-mail preservation, this XML schema is distinct in: 1) its account-based paradigm; 2) the granularity of data captured; 3) its alignment with the email message standard RFC 2822; 4) the support of a single XML file representation of the account; and 5) its incorporation into two separately developed e-mail preservation software applications.*

## Introduction

The Collaborative Electronic Records Project (CERP) originated in 2003 after a conversation between Dr. Edie Hedlin, Director of the Smithsonian Institution Archives, and Dr. Darwin H. Stapleton, Executive Director of the Rockefeller Archive Center (both since retired), about the state of electronic records. The Rockefeller Foundation partially funded the CERP grant proposal, and the Rockefeller University (at the time the RAC's parent institution) committed additional resources. In August 2005, each institution hired an archivist specifically for the project.

SIA is the institutional archives of the Smithsonian, being established by official directive in 1967. As part of its official role, it serves as the record manager of all units of the Institution. SIA collects, preserves, and makes available the official records of the Smithsonian Institution, the papers of Smithsonian scholars and other staff members, and the records of related professional organizations. It carries out a program of records management for Smithsonian offices, advising them on the disposition of records and pertinent documentary materials, and operates a Records Center for the temporary storage of scheduled records.

SIA has been accessioning born-digital records for more than a decade. In 2003, it established a formal Electronic Records Program to address growing digital curation and preservation needs. Email is transferred from a variety of systems, typically 5 years or more after becoming inactive.

The Smithsonian Institution's basic policy is to "create and keep complete and accurate records of its activities; maintain the integrity of those records; and preserve records of enduring evidential and historical value," according to *Smithsonian Directive 501, Archives and Records of the Smithsonian Institution.*

Both archivists conducted in-person interviews to assess depositors' business processes and electronic records practices. The RAC project archivist surveyed sixteen organizations (forty-six interviews) and the SIA project archivist surveyed three units (forty interviews).

Recordkeeping guidance was authored by both archives. The documents covered various topics such as how to weed out junk/personal material from email accounts; how to manage email and digital collections with reference to the Department of Defense 5015.2; and how to transfer email accounts to the archives. Both archives stressed the principle that content determines recordworthiness, not the format of the item/s.

Early in the project, CERP decided it would pursue e-mail archiving as accounts rather than as individual messages, chiefly because: 1) the sheer volume precludes using scarce resources to preserve each message and document the contextual relationships; and 2) the value of preserving email messages "in situ" resolved issues of original order and overall metadata management and documentation.

There are different models of record acquisition to consider: (1) incremental harvesting of active email from multiple users in a system or (2) grouped transfer of inactive email from multiple systems. The latter could be file(s) transferred as one data file containing email messages, their attachments, and their organization within the original account or groups of individual emails. Both acquisition models applied to CERP. In some cases, an institution will have no control over when an email account arrives at its door, nor the format or organization of said account.

## Workflow

Once accounts were selected for testing, CERP drafted workflow procedures that continued to evolve during the project. Much of the workflow involved manual processes. The steps were:

- Transfer of source (PST, MSG, GroupWise, etc)
- Document transfer and object metadata. Update metadata narrative throughout process
- Conduct virus scan
- Make backup copy
- Conduct preservation assessment, which includes extracting attachments and running format file (JHOVE/DROID) script on attachments to detect issues, reviewing account
- Start finding aid
- Convert source file to MBOX format
- Parse MBOX file and validate XML output (Parser output includes attachments, bad messages, and message summary)
- Create METS file (was used for DSpace ingest)
- Finalize metadata narrative and finding aid
- ZIP parser output
- Deposit into repository

CERP adopted the Open Archival Information System (OAIS) Reference Model, following the concepts of the Submission Information Package (SIP), the Archival Information Package (AIP), and the Dissemination Information Package (DIP) from the OAIS Information Model.

- The SIP contains the source email received from the depositor and initial metadata from the depositor and updated by the archivist.
- The AIP contains the source email, the administrative and descriptive metadata (narrative, METS), finding aid/s, MBOX files, email preservation XML file, parsed attachments, bad messages from parser, and parser subject-sender log.
- The DIP could be the entire package for viewing/downloading or specific email message/messages.

### *In a perfect world*

The account would be reviewed by the user to remove sensitive and non-essential messages before the transfer. The accession would include documentation from the user indicating the structure, dates, and other pertinent information about the account. The email account capture would involve a streamlined, error-free transfer of an account via a secure method, e.g. ftp. The transfer would be verified, the email account would be free of viruses, and backed up to a separate drive. Attachments would be reviewed and analyzed for obsolescence issues. Processing information would be added throughout the procedure and a finding aid and METS file would be automatically generated. If the account was in a proprietary format, then conversion to the MBOX (generic email format) would be conducted. The MBOX output would be parsed to create a valid XML file of the account free of bad ("illegal") messages. The complete package would be ready for deposit into a trusted digital repository.

## Real world challenges

Acquisitions of any type of digital records can be problematic. The challenges of actual transfer, assessment, and conversion of the test accounts at SIA during CERP are reviewed in more detail below.

### *Transferring the accounts*

In 2005, some Smithsonian offices were using Microsoft Outlook Exchange for email while remaining units were being moved from GroupWise. After reviewing the results of the interviews with the testbed staff, specific accounts were selected for transfer. The plan was to use Microsoft Exmerge for Outlook and Nexic Personal Discovery for GroupWise to capture copies of the email accounts for secure transfer [1]. SIA was to receive these copies of email messages and attachments (as a collection) while the originals would remain within the account holder's application.

This plan required coordination with a contact at OCIO (Office of the Chief Information Officer), SIA, and the testbed participants. This proved time consuming due to access issues, schedules, and other projects being tackled at the Institution and meant delayed transfers of test material.

At the beginning of Phase 2, SIA had only two email accounts for testing from one unit. One person was leaving the Institution, and SIA thought it was important to capture her email and other digital material before her departure. She was instructed to search specific keywords on her account and create a PST [2]. She had difficulty creating a PST file within her Outlook account and the messages were exported instead as separate MSG files via SIA's secure server. Since this office is located offsite, immediate technical assistance from SIA was not possible on the PST creation. The MSG files were converted into a PST with the program Aid4Mail so the archivist could review the entire account with its structure intact within Outlook. The other account was a PST file that was transferred via that unit's ftp server.

Parameters for the captures were based on date, such as messages prior to 2005, and specific subject subfolders when applicable in coordination with existing records series from unit records disposition schedules.

Once the Exmerge capture was finally scheduled, though, one office had converted from GroupWise to Outlook Exchange, which eliminated the need to use Nexic Personal Discovery and meant only PST files to transfer. The process was conducted by an OCIO staffer and the CERP project manager. The captures were problematic, as the email was either too recent and/or failed to include all the requested data such as the Sent Items folder. The process was not easily automated and one account took three to four hours to complete. Scheduling, staff departures, and other projects made it difficult to attempt additional Outlook transfers using Exmerge. Thus, it was decided it would be easier for the SIA project archivist and CERP project manager to conduct the captures on site at the testbeds of the remaining email accounts and transfer to SIA's server.

This method proved to be a better approach for SIA. The project manager and archivist controlled when the transfers would take place and assisted the account holders with the process. These transfers took 30-90 minutes to complete. Because one account was relatively small, an attempt at emailing the PST as an attachment to SIA was done. However, Outlook would not transmit the attachment because of SI's email security filters. Instead, a server transfer was conducted. It also was decided not to pursue email from some of the accounts that went through Exmerge initially because of time conflicts, employee schedules, and other projects.

### *Conducting preservation and managing sensitive content*

Ultimately, SIA captured eight email accounts for this pilot, totaling 2.7 GB or more than 36,000 email messages with attachments. There were more than 89,000 email messages for CERP.

Virus scans were conducted and backup copies were made of the testbed email accounts. Some accounts did contain viruses. Notifications were sent to those whose material was successfully transferred. A metadata narrative file was started at SIA indicating the collection name, method of transfer, size of account, number of messages, and other information. The file was updated throughout the processing of the account documenting tools used and conversion procedures taken.

The account holders were asked to weed their accounts of messages that should not be part of the test, such as personal and transitory messages, and follow-up email reminders were sent as the capture date neared. Some complied better than others. Non-business or non-essential emails remained in some accounts, though, such as news alerts from CNN, restaurant reservations, and school and church notifications.

Since SIA only had the two email accounts initially, there was time to explore them more fully on an item-level basis to review content, folder structures, and relationships as opposed to the later, and sometimes much larger, transfers. The archivist also reviewed sender information and subject lines. CERP was interested in the Internet Headers, as an authenticity marker [3]. Many were missing when viewed within Outlook at SIA due to migration from other email applications (GroupWise to Outlook Exchange) or because the messages were sent within the same mail server and failed to go through a SMTP server where Message IDs are added.

Keyword searching was conducted in these early transfers to test the practicability of this sorting/weeding method during processing. Relying on the search mechanism within Outlook was problematic as it lacked focus. A free unsupported application called Lookout (now part of Microsoft) provided better results. For example, using one account, the Outlook search "mission" had 128 hits. This included the terms "commission" and "submission." Lookout had 43 hits.

SIA**'s** record managers were consulted regarding the feasibility of using keyword searches for weeding purposes of email accounts when only an Inbox/Sent Items structure or other non-subject system was used. One such account contained more than 20,000 email messages. Some keywords were constructed from records disposition schedules or the information gathered from the testbed interviews. Ultimately, it was determined that recordworthy material could be missed using this method and that it would be a time-consuming exercise with larger accounts. Keywords also were not be used as parameters to capture email messages for the former reason.

Another example of why keyword searching could be problematic involved a video attachment. A review of some attachments within a 1.5 GB account revealed a non-business-related email from a colleague at another institution with a video of a skateboarding bulldog that has been featured on numerous websites and television. The recipient at SI was blind carbon copied. A few months later the recipient replied to that same email with a professional inquiry. She retained the original subject line, which had nothing to do with the business-related question. The respondent also kept that same subject line. This resulted in business and non-business messages being intermingled. If a researcher is looking for the business-related email message and only browsing/searching subject lines, it could be missed because it is labeled "skateboarding dog" and not "contract information."

Format identification of email attachments was an important issue due to the variety of file formats found in email attachments and their separate obsolescence factors. To prepare for this, the attachments were copied out of the email account in their native formats. Aid4Mail initially was used, but failed to retrieve attachments within child messages of messages. EZDetach from TechHit proved to be a more effective tool to use within Outlook (originals remain with source email). All extracted attachments were stored within their corresponding folders from the email account.

Once the attachments were extracted, file formats were analayzed using format identification tools JHOVE and DROID [4]. JHOVE provides robust metadata for a small set of standard-based file formats, while DROID handles a much larger range of formats. JHOVE required significantly more technical skills to install at SIA. This is offset by DROID's comparatively limited metadata output. Using both programs for assessments provide a good comparison mechanism and were adopted for the pilot. Outputs from both can be saved as XML.

Email attachments within a collection typically are not one format, as in the case where an archivist has image files saved as TIFFs and can use the TIFF module within JHOVE to get one report. Due to the multiple and proprietary formats within email collections, JHOVE presents limits in that regard. Obviously, the PDF module will report that there is a problem with a Microsoft Word document and a TIFF document. DROID, on the other hand, recognizes more than 100 formats, including Microsoft Office formats, but the metadata is extremely limited. DROID was a simple download and is also Java-based like JHOVE.

SIA developed a Java-based script that automates analyses of the attachments using both programs. The script generates: 1) a file log listing all the analyzed attachments; 2) a file list of the analyzed attachments and possible types determined by DROID and JHOVE for each; 3) outputs from the JHOVE modules and DROID; and 4) and a warnings file. This warnings file can contain the diagnosis from DROID when there is a possible file mismatch and JHOVE's analysis as well on that file in question. All output files can be reviewed to get a thorough analysis.

A primary goal of developing this script was to save format analysis time by eliminating the need to manually run the attachments through DROID and each JHOVE module separately. The warnings file serves only as a starting point to make the review of questionable files easier by logging results from both programs in a simple text document that an archivist can use to zero in on problematic files.

The team also grappled with the issue of these extracted native attachments. Should they be retained as part of the AIP? Should the base64 versions of the attachments from the parser be converted on the fly [5]? What about viruses within? A Windows check would fail to detect a rare virus for Mac and Linux. These questions were not fully answered during the project.

As SIA reviewed attachments, various issues arose: WordPerfect files with auto format for the date (which displays the date one is viewing the file rather than its real creation date); sensitive information such as Social Security numbers; broken animation files; duplicates; and renderability problems.

One account that was not part of the testbed sets was used for CERP demonstration purposes. Some weeding was performed on it due to the sensitive material including employee names and Social Security numbers contained within attachments without encryption. This processing was done manually in about 10 hours on 6,000 email messages. The original account was maintained.

## Preservation tool design and testing

For the pilot, SIA worked with PST files, which can only be opened in Outlook and can become corrupted around the 2 GB threshold. The format has already been altered by Microsoft, and it is possible PST could be eliminated. Other CERP testbed email formats included AppleMail, Eudora, GroupWise, and LotusNotes. These proprietary formats are not viable long-term preservation solutions.

After the IT consultant joined the project team, discussions focused on the need for a standard schema as a structure for preserved email accounts. Meanwhile, the National Historic Publications and Records Commission (NHPRC)-funded EMCAP project was also exploring email capture and preservation challenges [6]. CERP consultant Steve Burbeck and North Carolina State Archives technical contact David Minor began collaborating on the email account schema started by Minor (http://www.archives.ncdcr.gov/mail-account) that both projects are now using. While the E-Mail Account schema details were being refined and improved, the CERP consultant started developing a parser to create the XML output, resulting in a prototype built in an open source development system -- Squeak Smalltalk v3.9 (http://www.squeak.org). It can be run directly from the parser or a Web User Interface built with a popular Squeak Web Application development framework called Seaside (www.seaside.st).

The parser was designed to accept the MBOX format for processing. MBOX is a generic email format that offers a combination of openness and cross-platform support, unlike proprietary email formats. Most email clients can export mail in MBOX format and there are translation tools for converting various email formats to MBOX. It also makes it simpler for the parser to work with only one format. CERP initially used Aid4Mail from Fookes for the conversion of the PST into the generic format. While preparing an account for parser testing, SIA detected that some email message bodies were being separated as attachments when running through Aid4Mail. Email attachments also were missing or attachments were created such as winmail.dat files out of email bodies while another email had both its attachment and email message body missing prior to an upgrade to the software. Once the parsing started the consultant reported that the generic file from Aid4Mail was “close to MBOX format but not exactly” due to extra lines being added at the start of each email message. RAC reported that it did not have these issues with non-PST files when using Aid4Mail.

This led to more research into other conversion tools. SIA started testing MessageSave from TechHit, which works as an add-in with Outlook. According to the CERP consultant, the product handled Outlook idiosyncrasies well by creating complete MBOX files that are RFC 2822-compliant, resulting in better parser XML output of the email account [7]. SIA decided to use it for the conversion while RAC continued to use Aid4Mail for its non-PST email formats.

Initial testing was conducted on the consultant’s computer and the archivists were able to review the output from the parser for quality assurance and integrity. The parser generates a single file of the entire account rather than creating separate XML files for each email message. This approach means streamlined metadata management and produces preserved folder/message hierarchies. Any attachments larger than 25K are saved as separate XML encoded files. The attachment size threshold can be higher but CERP set this at 25K for data throughput purposes. Messages that are considered “bad” (malformed issues, illegal subject character lines, or unknown content types) by the parser also are output as single files so the archivist can view them individually. The last item is a spreadsheet that is useful as another access aid for archivists and researchers; referred to as the Subject-Sender log, it contains the message subject, sender, date, hash, and message ID.

After six months of code changes and tweaks, the parser was installed at SIA. Improvements continued to address issues such as modifying date format and accepting any MBOX file name (all files had to be named messages.mbox initially), along with the addition of the Web User Interface. Folders also had to be manually created by the SIA archivist for each MBOX file created from MessageSave in order to maintain the structure from the account. A script was written at SIA to create these folders at the various levels with their names and to place the MBOX file into its corresponding folder.

All of the SIA testbed accounts were parsed, and the email preservation XML files validated against the E-Mail Account schema. At this point, the XML output has to be manually checked against the PST to ensure integrity. Sampling is done with large accounts. Automation tools would be helpful with this step.

## Selecting XML for the preservation format

Using XML as the preservation format was appealing because it is open, human-readable and self-describing. Working with a schema, email accounts could be preserved in a consistent format that was both user-accessible and database-friendly. XML-preserved email messages could be presented in a user friendly

display while robust querying tools leverage a preserved account's tags to facilitate intensive research and data mining.

PDF and PDF/A formats were not chosen because their construction and capacity for content were ill-suited to capturing a full email message record: the highly structured, hidden content; the regular viewable content; and the attachments' content information. Aside from these limitations, selecting PDF or PDF/A as the preservation format would have precluded preserving an email account, with its folders and messages intact, as a single file, and thus would have required additional metadata to be created in order to relate the individual messages, perhaps tens of thousands in a single account, to each other.

## The E-mail Account Schema

Unlike other work in the area of e-mail preservation, the E-mail Account schema is distinct in: 1) its account-based paradigm; 2) the granularity of data captured; 3) its alignment with the email message standard RFC 2822; 4) the support of a single XML file representation of the account; and 5) its incorporation into two separately developed e-mail preservation software applications. The E-Mail Account schema has made possible the preservation of large bodies of related e-mail in a single XML file.

### *Earlier email preservation efforts*

Early work on email preservation recognized XML as a preferred preservation file format. Three notable efforts developed effective solutions preserving an email message in an XML file.

The Dutch National Archive's Digital Preservation Testbed (DPT) included email in its work on digital preservation strategies for typical office records. Focused on retaining individual email messages, it developed an preservation tool that works as a helper application for Microsoft Outlook. The tool, XMail, used a project-developed XML schema to migrate significant values of the message into an XML file [8]. The DPT recommendation for email preservation was published in the report series "From Digital Transience to Digital Durability" in 2003 [9].

The four-year DAVID project also looked at email as it worked to address archival and legal concerns [10]. It, too, developed a preservation tool that incorporated a project-defined XML DTD that addressed individual messages, migrating each message into an XML format.

The National Archives of Australia's XENA preservation software works on email messages or mail datastores. It determines the preservation format based on the email message's file format. An older PST file would be broken into individual messages, and then preserved as XML files and a related index file; however, an HTML-formatted email message would be converted to XHTML. (Note: Both SIA and RAC were unable to convert PST files using XENA. Online references indicated XENA does not work with Outlook 2003 currently [11].)

In each of these cases, the end result was an individually preserved email message. Research done across groups of email messages would require additional effort on the part of the researcher to reconstruct the relationships that had been in place prior to preservation.

### *Collaboration*

Seeking to retain the metadata inherent in an email account and its presentation of email messages, CERP and the EMCAP project worked together to define an XML schema that effectively captured and preserved email messages 'in situ' in such a way that they retained full authenticity and integrity of each message while enabling researchers to use robust search and data mining strategies to identify valuable content in individual messages, within folders or accounts, and hopefully across accounts.

The collaboration yielded the E-Mail Account schema which accomplishes these goals. The schema has been implemented in the email preservation tools of both projects. The tools are written for different acquisitions models, and the schema proves effective under both scenarios.

### *Details and Structure*

The schema leverages XML's nested tagging structure to embed the organization and structure inherent in an email account. Beyond the most basic organizational structure of an email account with a folder that contains at least one message, the schema needed to be robust enough to handle multi-format messages, messages with attachments, and messages with attached messages while at the same time capturing the multi-tiered structural organization given to the email account by the account owner. The fully developed schema provides that capability incorporating the account organization into a single XML file for the whole account and its messages [12].

### *Folders: self-describing organization*

A certain amount of structured organization is predefined by the email system supporting an account. When an email account owner expands this predefined structure by adding additional folders and subfolders during use of the account, these document relationships imposed by the account owner on groups of messages, becoming valuable metadata helpful to future researchers trying to grasp the significance of email within the larger body of the account.

The E-mail Account Schema structure presents the email messages in the folders that contain them. If the account owner had developed a multi-level organization scheme, the schema presents these as <folder> tags nested within <folder> tags until the full hierarchy has been described. Just as an active folder can contain both messages and other folders, an account preserved with the schema supports both as this is a common occurrence in email accounts. The structure of a preserved account file is partially illustrated below.

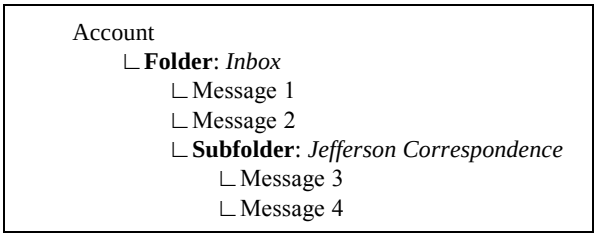


***Figure 1**. Partial structure of a preserved email account.*

### *Messages: simple and complex*

The E-mail Account Schema carries forward the elements defined in the RFC 2822 for account messages. Therefore, an email message's preserved components extend beyond the limited set of elements viewable by a typical email user. A short list of message header elements: LocalId, MessageId, MimeVersion, OrigDate, From, Sender, To, Cc, Bcc, InReplyTo, References, Subject, Comments, Keywords, demonstrates how the association of the schema with the RFC 2822 standard works to ensure a full representation of the original email message's content information can be preserved in the output.

Most popular email systems are capable of generating multi-body types; messages that include either or both HTML and plain text, leaving it to the recipient's email client to select which body type to display.

The schema supports embedding email message attachments in the preserved account file. When this occurs the embedded attachment is kept within the message. Alternatively, the schema allows for a message attachment to be archived external to the email account XML file, specifying how this is documented in the preserved account file.

The end result is that a thorough preservation of an email message in its entirety – headers, message, and attachments – is accomplished. Whether a plain text email without attachments, or a multi-body email with documents, images, videos, and other emails attached, this range of possibilities is handled in the schema's definition.

## Potential values of the E-mail Account Schema

The E-mail Account schema is distinct in: 1) its account-based paradigm; 2) the granularity of data captured; 3) its alignment with the email message standard RFC 2822; 4) the support of a single XML file representation of the account; and 5) its incorporation into two separately developed e-mail preservation software applications.

A key value of the account-based paradigm is that the interrelationships of the email messages themselves are preserved without requiring additional documentation as the information already exists within the account. The burden of metadata management is therefore reduced because it remains with the archived messages. The original order, the thread index values, etc. are preserved right along with the email body content.

The schema itself serves as a means of validating that a preservation migration was completed successfully. When accounts contain tens of thousands of emails, an efficient means of verifying the quality of completed preservation processes is essential. Similarly, it could be used in a digital object repository as a means of confirming whether a digital object presenting itself to the repository as a preserved email account is an email account.

The adherence to RFC 2822 provides a more comprehensive and complete range of data, organized in a standard-based format that makes it more accessible. It also introduces the opportunity for email system vendors to adopt the schema as a data output option, facilitating the future archiving of email accounts. The schema can also be incorporated into other email preservation software.

The granularity of the schema structure facilitates the accessibility and understandability of preserved email accounts and their messages by enabling advanced searching strategies to be applied to one or more accounts simultaneously. This helps to expose social networks and message interrelationships present in, and across, accounts.

Because of the schema's organization, it is possible to search throughout an account, then return only those messages that satisfied the criteria for display to the user. This may possibly be extended to cross-account result sets.

In discussions with other archivists, the potential for facilitating research of social networks as documented in emails has been particularly noted. These networks can be exposed by querying and mapping message header elements. With a consistent structure between preserved accounts, these searches could be conducted across multiple accounts and only those elements that meet the criteria be returned to the searcher for viewing.

These represent a few of the values that an account-based paradigm for email preservation, and the E-Mail Account schema hold for digital curators and archivists.

## References


[1] Exmerge is a MS Exchange utility program that extracts data from mailboxes on an Exchange Server; and Nexic Personal Discovery allows the export of messages from an account in ASCII text format.

[2] PST stands for Personal Storage or Personal Stores within Outlook. It stores email and attachments outside of the email server in a single file that can be saved on a network server, a hard drive, or removable media. One can view all messages and attachments in a PST file within Outlook.

[3] An Internet Header contains sender and recipient IP addresses, domain names, times, and Message IDs. It does not normally display in an email client and typically has to be opened in a separate step.

[4] JHOVE is the JSTOR/Harvard Object Validation Environment. DROID is the Digital Record Object Identification from the National Archives in the United Kingdom.

[5] Base64 is a binary-to-text encoding schema. Others include hexadecimal, quoted-printable, and BinHex.

[6] EMCAP is the Electronic Mail Capture and Preservation project, which is being conducted jointly among North Carolina, Pennsylvania, and Kentucky.

[7] RFC 2822 is the Internet Message Format. The "standard specifies a syntax for text messages that are sent between computer users, within the framework of 'electronic mail.' " -- Available at http://www.w3.org/Protocols/rfc822/. Accessed Dec 1, 2008.

[8] XMaiL software: http://www.digitaleduurzaamheid.nl/index.cfm?paginakeuze=299_

[9] © Digital Preservation Testbed, The Hague, 2003. Digital Preservation Testbed: From digital volatility to digital permanence. Preserving E-mail (2003)

[10] DAVID and eDavid project website: http://www.expertisecentrumdavid.be

[11] Available online at http://sourceforge.net/tracker/index.php?func=detail&aid=1946019&group_id=85722&atid=577089. Accessed Dec 1, 2008.

[12] RDL documentation of the E-Mail Account schema. http://www.archives.ncdcr.gov/mail-account


## Author Biography

*Riccardo Ferrante received his BS in education and social policy from Northwestern University. He has led the Electronic Records Program at the Smithsonian Institution Archives since its inception in 2003. His work in digital preservation, curation, and digitization includes a focus on born-digital objects and trustworthy repositories. He is a member of the Society of American Archivists.*

*Lynda Schmitz Fuhrig received her BS in print communications from Bradley University and her MA in history from the University of Illinois at Springfield. She has worked at the Smithsonian Institution Archives-Electronic Records Division since 2005. Her work has focused on email preservation, digitization, and other digital curation/preservation issues. She is a member of Society of American Archivists and Mid-Atlantic Regional Archives Conference.*