THE ROCKEFELLER ARCHIVE CENTER

Smithsonian Institution Archives

Collaborative Electronic Records Project

From Here to Eternity (Or Close To It): Phases 2 and 3 of CERP at the Smithsonian Institution Archives

Table of Contents

This documentation is released by the Collaborative Electronic Records Project under a Creative Commons Attribution-Non-commercial-Share Alike 3.0 United States License, 2008 and 2009. This license can be viewed at http://creativecommons.org/licenses/by-nc-sa/3.0/us/

Citation
Schmitz Fuhrig, Lynda. (2008). "From Here to Eternity (Or Close To It): Phases 2 and 3 of CERP at the Smithsonian Institution Archives." Sleepy Hollow, NY and Washington DC: The Collaborative Electronic Records Project.

The Rockefeller Archive Center (RAC) and the Smithsonian Institution Archives (SIA) launched the three-year, grant-funded Collaborative Electronic Records Project (CERP) in August 2005. The project soon narrowed its focus from born-digital material to email processing and preservation due to the significance that this communication form plays in everyday business operations among depositing offices and organizations. While both institutions fulfill archival missions, their operations are quite different. For example, SIA is the record manager for Smithsonian Institution while the RAC has no control over its depositors. For additional background on the development of CERP, see *"Collaborative Electronic Records Project: An Introduction and Overview."*

As CERP completed Phase 1 tasks of researching digital preservation issues, interviewing testbed participants, and compiling those results, the focus became more technical. The next two phases involved transferring and processing email accounts, testing tools, and devising and refining functional and system requirements and products such as the preservation parser. This paper outlines SIA's experience during these phases.

Transfer of email accounts

Typically, the email collections were inactive for CERP and that is the case at SIA outside of this pilot. For SIA testbeds, two accounts were inactive and one was about to become inactive. The remaining accounts were still active, in that the users still had the messages and attachments in their active Outlook accounts. In all cases only copies of the accounts were transferred to SIA as PST files.[1] The SI email systems are completely separate from the CERP server and repository.

In 2005, some Smithsonian offices were using Microsoft Outlook Exchange for email while remaining units were being moved from GroupWise. 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.[2] 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

[1] PST stands for Personal Storage or Personal Stores within Outlook. The PST stores email and attachments outside of the email server as one file of all the email messages and attachments saved to it.
[2] Exmerge is a MS Exchange utility program that can extract data from mailboxes on an Exchange Server; and Nexic Personal Discovery allows the export of messages from an account in ASCII text format.

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. 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.[3] 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.

SIA created transfer documentation for the participating testbeds. The guidelines indicated why the account was selected, i.e., email considered recordworthy and not recent, how the files would be transferred, processed, and stored, and outlined post-project procedures. 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.

SIA issued email guidance defining what makes email a record, tips for weeding email accounts, and some consequences of poor email management. SIA also offered assistance to email users on search capabilities and PST creation. Participants were told that access to the test files would be limited only to the SIA CERP team and the SIA Records Manager liaison.

An SIA chart of recommended digital preservation formats also was available to the testbeds, as well as to the entire Institution.

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

[3] Smithsonian Institution follows strict computer technology protocols as defined in various federal guidelines and best practices.

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.

SIA also transferred other digital files from the participating testbeds that were possibly recordworthy for permanent accessioning into the Archives. These files included a unit handbook, digitized historical documents, and various reports. This was accomplished on site by transferring files to SIA's secure server or by allowing the unit access to place documents on SIA's server. This latter method had mixed results. The files transferred correctly, but the connection to the server failed on a regular basis when the user tried to access it.

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. Overall, there were more than 89,000 email messages for the two archives.

Virus scans were conducted and backup copies were made of the testbed email accounts. 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 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. Both SIA and RAC also were interested in the Internet Headers, as an authenticity marker.[4] Many were missing when viewed within Outlook at SIA. The missing Internet Headers were 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.

SIA conducted some keyword searching 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.

The project archivist and project manager also consulted with the SIA Records Management Team about 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. 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 an 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

[4] An Internet Header contains the sender and recipient's IP addresses, domain names, times, and Message IDs. It is not part of the body of the email that displays in an email client and typically has to be opened in a separate step.

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."

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. This processing was done manually in about 10 hours on 6,000 email messages. The original account was maintained.

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.

When the SIA project archivist opened a MS Word document on her PC, the file appeared in an unreadable font (similar to Dingbats - ♎♓■♑♌♋♦⬧). After changing the typeface to another font, the display remained problematic. Opened in OpenOffice, which is open-source office suite software, the file was legible. After viewing the document in NotePad to find out about the Word fonts, it was determined the format conversion was set on the archivist's PC to ESRI fonts, which are from the GIS software. [5] ArcGIS was installed on the PC in 2006, and the archivist was not aware of the font replacement change. There was an attachment within the attachment too.

The challenges before CERP were preserving the email accounts in a reliable, sustainable digital format and using a trustworthy repository to store them. 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. *See Appendix 1.*

[5] PCs running Windows/Microsoft Office allow for the automatic substitution of another font within a document for the one not installed on the machine.

A PST can only be opened in Outlook and PST files can become corrupted around the 2 GB threshold. The format has already been altered by Microsoft, and it is possible PST could be eliminated. This proprietary format is not a viable long-term preservation solution.

The CERP team discussed the option of using XML as the preservation format for the email collections. XML was appealing because it is open, human-readable and self-describing. With the right web translator, it can be presented in a user friendly display.[6] Other institutions such as the Antwerp City Archives and National Archive of the Netherlands used XML as well for their email projects. While these projects focused on individual messages preserved in XML, CERP decided to approach email preservation on the account level, which maintains the structure and relationships within that collection and simplifies metadata management. XENA software from the National Archives of Australia also relies on XML. Both SIA and RAC were unable to convert PST files using Xena though. Online references indicated Xena does not work with Outlook 2003 currently.[7]

PDF and PDF/A were not chosen because of their limitations. While the format can replicate the on-screen appearance of an email message, attachments fail to transfer with some conversion programs and Internet Header information and attachment relationships can be difficult to capture.

Once the IT consultant joined the project team, discussions focused on the need for a standard schema.[8] Meanwhile, the National Historic Publications and Records Commission (NHPRC)-funded EMCAP project was also exploring email capture and preservation challenges.[9] 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

[6] XML is currently used in many websites and its web page display is achieved through XSLT, CSS, and/or Javascript.
[7] Available online at http://sourceforge.net/tracker/index.php?func=detail&aid=1946019&group_id=85722&atid=577089. Accessed Dec 1, 2008.
[8] See http://www.w3schools.com/Schema/default.asp for more on schema.
[9] EMCAP is the Electronic Mail Capture and Preservation project, which is being conducted jointly among North Carolina, Pennsylvania, and Kentucky.

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. SIA initially used Aid4Mail from Fookes for the conversion of the PST into the generic format. While preparing an account for parser testing, SIA noticed 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.[10] 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 throughput and to keep the main XML file a manageable size. 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

[10] 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.

archivist can view them individually. The last item is a spreadsheet that was initially intended only for the consultant but is useful as another access aid for archivists and researchers, as it contains the message subject, sender, date, hash, and message ID. It is known as the Subject-Sender log. The SIA archivist previously had tried various ways of producing a spreadsheet of email addresses from the accounts.

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. Speed varied on PCs depending on machine specifications, but a coding change did improve processing time. 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. For example, an Inbox folder with 89 subfolders meant 90 folders constructed by hand. 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.

Format identification of email attachments was an important issue as part of SIA's best practices due to the variety of file formats found in email collections and their separate obsolescence factors. This required an extraction of the attachments in their native formats. Aid4Mail initially was used, but the software only captured the first level, failing to retrieve attachments within child messages of messages. EZDetach from TechHit proved to be a more thorough tool to use within Outlook (originals remain with source email). All extracted attachments are stored within their corresponding folders from the email account.

Once the attachments were extracted, the file format identification tools JHOVE and DROID were applied to the collections.[11] JHOVE provides robust metadata for a small set of standard-based file formats, while DROID handles a much

[11] JHOVE is the JSTOR/Harvard Object Validation Environment. JHOVE2 is in the works; and DROID is the Digital Record Object Identification from the National Archives in the United Kingdom.

larger range of formats. JHOVE required significantly more technical skills to install at SIA.[12] 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?[13] 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.

[12] Troubleshooting was required due to java and configuration file issues on the SIA workstation. The Harvard team was very helpful.
[13] Base64 is a binary-to-text encoding schema. Others include hexadecimal, quoted-printable, and BinHex.

CERP also wanted an online finding aid as part of the package. SIA decided it wanted to explore Encoded Archival Description (EAD) for the pilot after deciding in 2001 it was not the right fit for the Archives. SIA's current finding aids are not as extensive as what EAD offers. The SIA project archivist and project manager met with the Archives of American Art at the Smithsonian to discuss its EAD workflow and standards. The project archivist also researched other repositories' use and creation of EAD files, reviewed the EAD Cookbook, and took an online course. The archivist and project manager also met with the SIA Records Management Team to discuss its finding aid workflow and tools used. It is something SIA may revisit after the project's completion.

A general digital workflow document specific to the email project was prepared. The stages are: 1) transferring the account and its metadata (the SIP); 2) processing and analyzing the account, virus scanning, transforming to MBOX, and parsing; 3) creating the METS files (there are two: one is the ingest mechanism for the digital repository and a copy serves as the traditional METS wrapper); and 4) zipping the entire package (AIP) for a digital repository deposit. *See Appendix 2.*

- 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.

The AIPs needed a digital repository. The team explored Fedora and DSpace, eventually settling on DSpace due to its maturity, strong user community, and its use already at the Smithsonian Institution Libraries and Rockefeller University (at that time, the RAC's parent organization). Both Fedora and DSpace require technical skills and customization for installation. DSpace also cannot handle hierarchical relationships, thus requiring the zipping of various components (finding aids, parser output, etc.) in the AIP.

CERP used a technology adviser with expertise in DSpace. Already working with Rockefeller University's DSpace instance, Lawry Persaud focused on the challenge of importing an AIP, particularly but not limited to an email account

AIP through a METS document.[14] This was something that was not achieved previously.

The DSpace Ingest Package Plugin now uses the METS document to conduct the ingest of the AIP. This was a natural step to pursue since CERP was already using a METS document as the metadata wrapper for the AIP. This METS import file contains the multiple names of those email AIP elements and describes the MIMETYPE, ID, size, and location. Once in DSpace, the email collection displays all the AIP files associated with the item. *See Figure 1.*

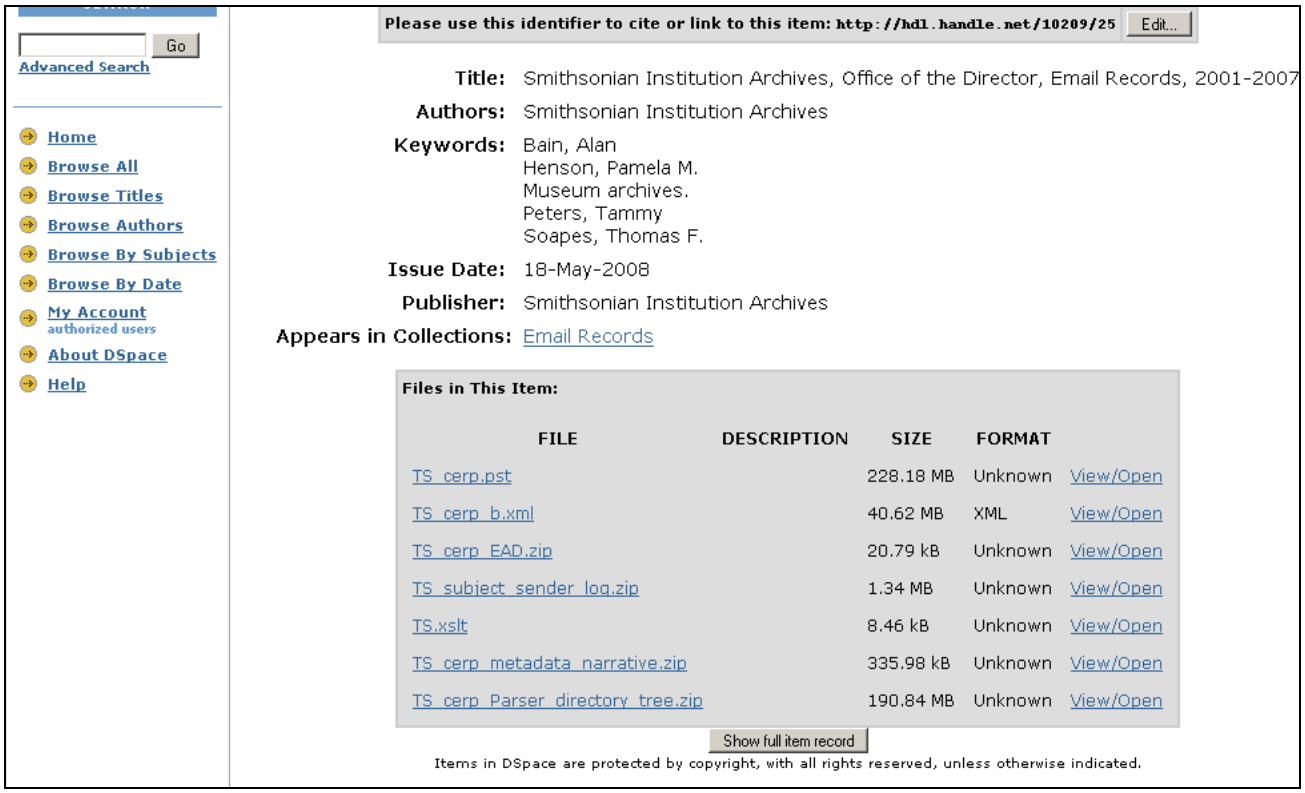


Figure 1

Metadata tags were needed for the METS document. The CERP team selected the Dublin Core metadata element set since DSpace uses this standard. The team settled on fields that seemed most appropriate for the METS ingest file, which uses the metsHdr (METS Header), dmdSec (Descriptive Metadata Section), the FileSec (File Section), and StructMap (Structure Map). A METS file must include these sections in order to comply with the METS standard.

The METS file was structured in the following way:

[14] See http://www.loc.gov/standards/mets for more information about this metadata schema.

The **METS Header** identifies the file as XML in syntax and follows the METS XML schema. It lists properties of the METS document such as the author (the person who created the METS doc), createdate, lastmodifieddate, and note.

Descriptive Metadata Section

For the purposes of CERP, the 10 Dublin Core elements below are mandatory because they are used as search and indexing criteria by the DSpace repository. However, the DmdSec can contain many more descriptors whether they are Dublin Core, other data standards, or custom fields.

<dc:publisher>SIA</dc:publisher>

<dc:relation.ispartofpublisher>Office of the Director</dc:relation.ispartofpublisher>

<dc:relation.ispartofpublisher>Email Records</dc:relation.ispartofpublisher>

<dc:creator>Soapes, Thomas F.</dc:creator>

<dc:contributor>Smithsonian Institution Archives</dc:contributor>

<dc:identifier.other>Accession 07-109</dc:identifier.other>

<dc:title>Smithsonian Institution Archives, Office of the Director, Email Records, 2001-2007</dc:title>

<dc:date>2001-2007</dc:date>

<dc:description.tableofcontents>Alan's reports</dc:description.tableofcontents>[15]
<dc:description.tableofcontents>CERP--PROJECT</dc:description.tableofcontents>
<dc:description.tableofcontents>Pam's reports </dc:description.tableofcontents>
<dc:description.tableofcontents>Sent Items </dc:description.tableofcontents>
<dc:description.tableofcontents>SIA Move </dc:description.tableofcontents>
<dc:description.tableofcontents>Tammy's reports</dc:description.tableofcontents>

<dc:subject>Bain, Alan</dc:subject>
<dc:subject>Henson, Pamela M.</dc:subject>
<dc:subject> Museum archives. </dc:subject>
<dc:subject>Peters, Tammy </dc:subject>
<dc:subject>Soapes, Thomas F. </dc:subject>

File Section

For each file described by this METS file, by FileGroup <FileGrp>.
ID

[15] Folder/subfolders within an email collection populate the <dc:description.tableofcontents> tag.

MIMETYPE
SIZE
LOCATION

Structural Map
A map of all files in the AIP by their relative locations. Listed by the FileID established in the above section

The other METS sections were not used. See *Appendix 3* for a diagram of how the XML schema follows the account structure and the full METS file of an email collection.

The team also worked on issues of metadata mapping within DSpace, Dublin Core, and EAD. Specific terms led to confusion within the various applications, such as author. In DSpace the author label applies to author of the digital item or Title within the repository. In EAD author applies to the author of the finding aid. In Dublin Core there is no dc:author field, only dc:creator or dc:contributor.author; we used dc:creator for the name of the email account holder.

The team discussed homepage design, security, and access issues within DSpace. The consultant built the site specifically for the pilot and led the installation of the server at RAC. While SIA was hoping for a search index built on the To, From, Date, and Subject lines fields within the top-level emails in DSpace, the scalability and resources were not available during the remaining time in the pilot. DSpace instead delivers the ability to search on the owner of the account and the descriptive metadata supplied by the Dublin Core tags listed above.

In addition, the researcher can search subject lines within the subject-sender log. Another goal is to have the email account in XML display as HTML in a browser. The team also talked about possibility of full-text indexing on the subject-sender log. Some basic searching development was started late in the project based on the content from within the parsing tool.

The future
The CERP group achieved its goal of preserving email for the long term in a sustainable format and storing those collections in a digital environment. The

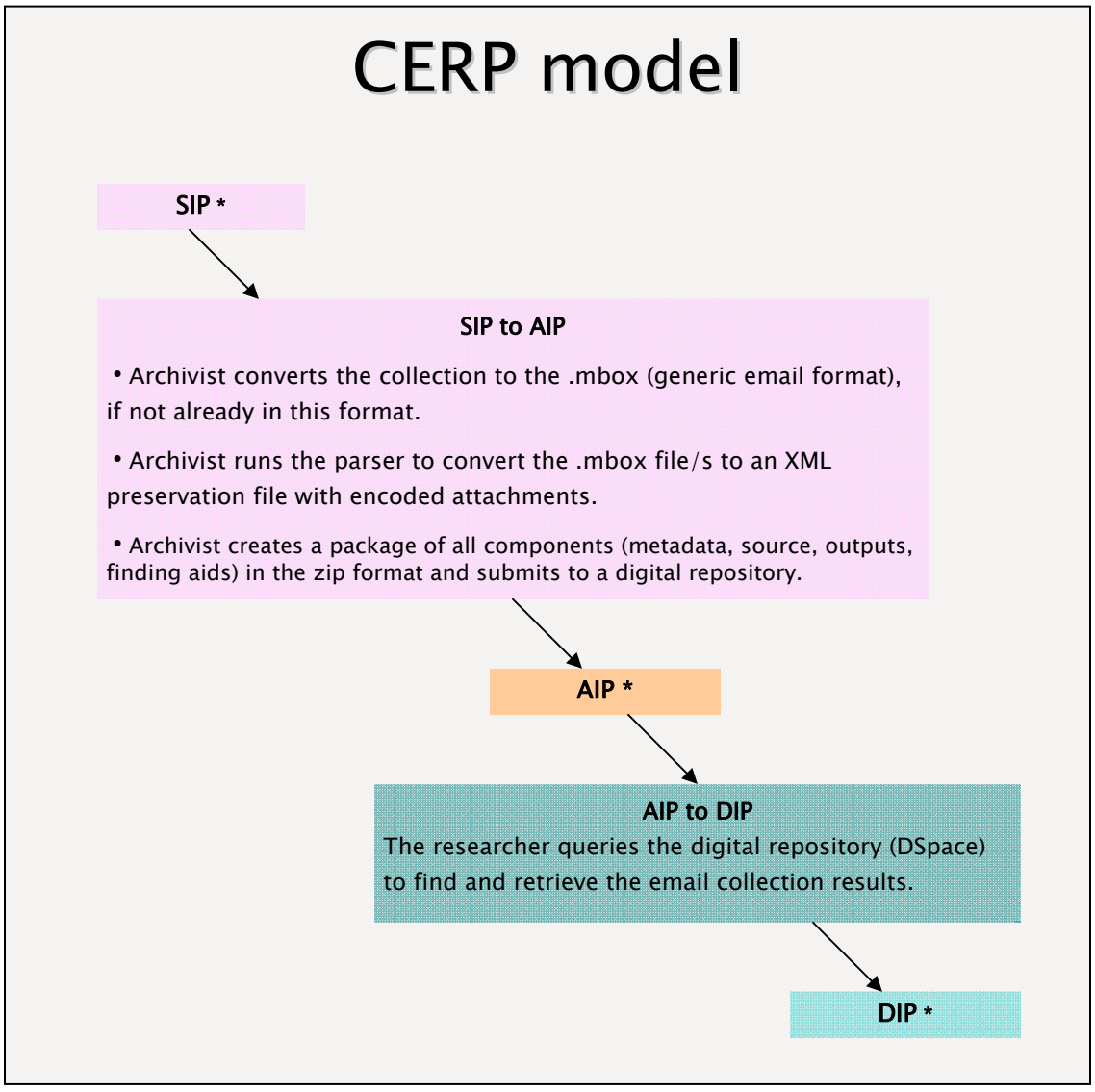
lessons learned have been shared with the greater archival community through the CERP website, numerous presentations, and newsletters.[16]

Those wanting to carry this work further should explore refined search and display, automated tools for weeding and verifying content, and social relationship/discovery tools.

Both CERP and EMCAP have proven that the XML schema results in a format that different organizations can use. The two projects would like to see other institutions adopt this schema in their email preservation work. It might lead developers of email applications to incorporate this preservation format into their products, contributing to future sustainability of important digital records.

[16] The CERP website will continue to be available indefinitely after the project's conclusion at http://www.siarchives.si.edu/cerp.

Appendix 1



** The SIP is the submission information package. It contains the email collection (variety of formats possible) received from the depositor and metadata narrative (both information supplied by the depositor and updated by the archivist).*

** The AIP is the archival information package. It contains the source email from the depositor, metadata (manually created METS, narrative, and other), finding aid (manually created), .mbox files, parsed XML file, parsed attachments, bad messages from parser, and parser subject-sender log.*

** The DIP is the dissemination information package. Package could include the entire package for viewing/downloading or a specific email message/s for viewing. The AIP remains in its original form.*

Appendix 2

Email Preservation Workflow

Stage	Components	What we do	Method
TRANSFER	Source Email	Receive	
	Metadata – Administrative & Descriptive	Document transfer and object metadata; make back up copy	Manual
SIP	*At this point we have a fully defined SIP*		
	Preservation/Conservation	ASSESS/RISK ANALYSIS: what formats, what media, obsolescence risks, viruses	Manual - Virus detection: AVG, Symantec, etc.; - Attachment extraction: EZDetach Partially automated - Attachment format identification: JHOVE, DROID
	Metadata – Administrative	Formalize assessment results	Manual
		Optional – Process to produce specialized DIP with sensitive data redacted *(*highly desirable but not intrinsically necessary for AIP)	Manual
	Metadata – Descriptive	Draft narrative finding aid	Manual
	Preservation	Complete preliminary transformation of grouped messages to MBOX format	Manual MessageSave, Aid4Mail
	Account.XML	PARSE (automated); validate	Automated CERP parser; oXygen XML Editor
	Preservation – MBOX Directory Tree (Parser Directory Tree)	Group attachments, bad messages, and message summary files within the existing MBOX directory tree and zip	Automatic from parser Manual zip
	Metadata, Descriptive – Directory tree of subject-sender logs	Group subject-sender logs (index) and zip	Automatic from parser Manual zip
	Metadata – Administrative	Finalize metadata including Preservation Description Information (PDI)	Manual
	METS	Describe AIP in METS schema and generate METS file. Duplicate METS file for loading into DSpace	Manual
	Metadata – Descriptive	Produce completed finding aid	Manual
	XSLT – display stylesheet	Include display template for access to preserved account	Manual
AIP	*AIP = all components above*		

TRANSFER

Source email
Basic accessioning metadata

SIP

Data Management
- Initial processing
- Backup
- Virus check
- Preservation assessment (physical/file/technical via JHOVE, DROID)

Preservation/Conservation
- Preliminary transformation – source format to MBOX format
- Final transformation (parsed) – MBOX format to XML

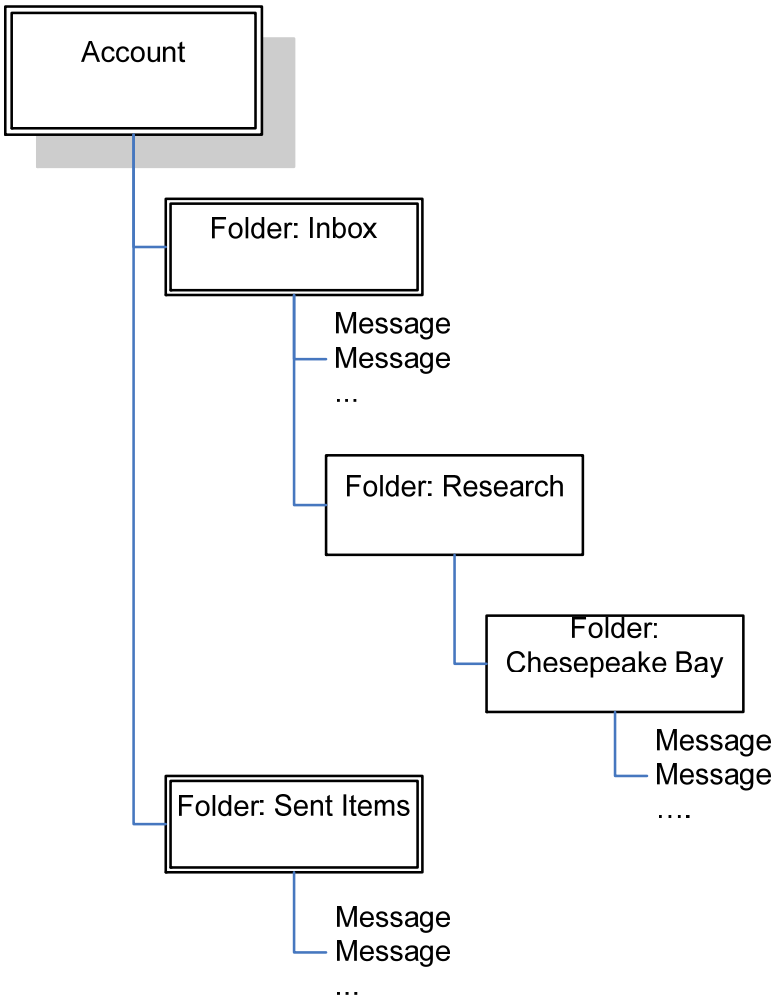
Metadata – Initial
- Metadata – administrative with preservation assessment
- Metadata – descriptive including narrative Finding Aid

AIP Assembly
- Source email (.pst, .msg, etc.)
- Preservation email
  - Parsed messages (XML)
- Parser Directory Tree.zip
  - Pre-parsed messages (MBOX)
  - Bad messages
  - Message Summary
  - Attachments
- Subject-sender log.zip (from parser output)
- Metadata narrative.zip – complete (administrative & descriptive)
- Finding aid narrative.zip
- METS.xml
- File Name METS.xml (administrative & descriptive metadata encoded in METS)
- Display aids
  - XML stylesheet

AIP

Appendix 3

The schema: email organization



SIA Sample METS file for email account

```
<?xml version="1.0" encoding="UTF-8"?>
<METS:mets ID="mets_1" OBJID="hdl:10209/150" LABEL="DSpace Item"
PROFILE="DSpace METS SIP Profile 1.0" xmlns:METS="http://www.loc.gov/METS/"
xmlns:xlink="http://www.w3.org/1999/xlink"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.loc.gov/METS/
http://www.loc.gov/standards/mets/mets.xsd">
  <METS:metsHdr  ID="H1" CREATEDATE="2008-02-08T06:32:00"
LASTMODDATE="2008-02-08T06:32:00" RECORDSTATUS="A">
    <METS:agent ID="A1" ROLE="CREATOR" TYPE="INDIVIDUAL">
      <METS:name>Schmitz Fuhrig, Lynda</METS:name>
      <METS:note>Smithsonian Institution Archives, Office of the Director, Email
Records, 2001-2007</METS:note>
    </METS:agent>
    <METS:agent ID="A2" ROLE="CUSTODIAN" TYPE="ORGANIZATION">
      <METS:name>SIA CERP </METS:name>
      <METS:note>Object owned by SIA </METS:note>
    </METS:agent>
  </METS:metsHdr>
  <METS:dmdSec ID="dmd_1" STATUS="A">
    <METS:mdWrap MIMETYPE="text/xml" MDTYPE="DC" LABEL="Dublin Core
Metadata">
      <METS:xmlData>
        <oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
xmlns:dc="http://purl.org/dc/elements/1.1/">
          <!--DSpace identifier-->   <dc:creator>Soapes, Thomas F.</dc:creator>
          <!--DSpace identifier-->   <dc:contributor>Smithsonian Institution
Archives</dc:contributor>
          <!--DSpace identifier-->   <dc:title>Smithsonian Institution Archives, Office
of the Director, Email Records, 2001-2007</dc:title>
          <!--DSpace identifier--> <dc:publisher>Smithsonian Institution Archives
</dc:publisher>
          <!--DSpace identifier--> <dc:relation.ispartofpublisher>Office of the
Director</dc:relation.ispartofpublisher>
          <!--DSpace identifier--> <dc:relation.ispartofpublisher>Email
Records</dc:relation.ispartofpublisher>
          <dc:dateaccessioned>February 8 2008</dc:dateaccessioned>
          <dc:date.created>April  2 2007</dc:date.created>
          <!--DSpace identifier-->   <dc:date>2001-2007</dc:date>
          <!--DSpace identifier-->   <dc:subject>Bain, Alan</dc:subject>
          <!--DSpace identifier-->     <dc:subject>Henson, Pamela M.</dc:subject>
          <!--DSpace identifier-->   <dc:subject> Museum archives.  </dc:subject>
```

<!--DSpace identifier--> <dc:subject>Peters, Tammy </dc:subject>
<!--DSpace identifier--> <dc:subject>Soapes, Thomas F. </dc:subject>
<dc.identifer.uri></dc.identifer.uri>
<dc:language>English</dc:language>
<dc:description.abstract>This accession consists of records created by Thomas F. Soapes during his tenure as Acting Director of the Smithsonian Institution Archives (2005-2007). It includes emailed reports from Pam Henson (Institutional History Division), Alan Bain (Technical Services Division), and Tammy Peters (Archives Division); CERP Project email; email related to the Archives' move from the Arts and Industries Building to Capital Gallery; sent email; and weekly manager reports (MS Word) sent to Shelia Burke (Deputy Secretary and Chief Operating Officer). Sent email also includes correspondence while he was chair of the Archives Division at the National Air and Space Museum. Some correspondence is transitory and/or sensitive in nature. </dc:description.abstract>
<!--DSpace identifier--><dc:identifier.other>Accession 07-109</dc:identifier.other>
<dc:description.note></dc:description.note>
<!--DSpace identifier--> <dc:description.tableofcontents>Alan's reports</dc:description.tableofcontents>
<!--DSpace identifier--> <dc:description.tableofcontents>CERP--PROJECT</dc:description.tableofcontents>
<!--DSpace identifier--> <dc:description.tableofcontents>Pam's reports </dc:description.tableofcontents>
<!--DSpace identifier--> <dc:description.tableofcontents>Sent Items </dc:description.tableofcontents>
<!--DSpace identifier--> <dc:description.tableofcontents>SIA Move </dc:description.tableofcontents>
<!--DSpace identifier--> <dc:description.tableofcontents>Tammy's reports</dc:description.tableofcontents>
<dc:rights></dc:rights>
<dc:accessrights>Unrestricted</dc:accessrights>
<dc:available>2008</dc:available>
<dc:type>Electronic mail</dc:type>
<dc:type>Mixed material</dc:type>
<dc:format.extent>210 MB</dc:format.extent>
<dc:format.extent>228 MB</dc:format.extent>
<dc:format.extent>40 MB</dc:format.extent>
<dc:format.medium>pst</dc:format.medium>
<dc:format.medium>mbox</dc:format.medium>
<dc:format.medium>xml</dc:format.medium>
<dc:source>SIA</dc:source>
<dc:relation></dc:relation>
<dc:coverage.temporal></dc:coverage.temporal>
<dc:coverage></dc:coverage>

```
            </oai_dc:dc>
          </METS:xmlData>
        </METS:mdWrap>
      </METS:dmdSec>
      <METS:fileSec>
        <METS:fileGrp USE="CONTENT">
          <METS:file ID="TS_TEST_1" MIMETYPE="application/xml" SEQ="1" SIZE="40000"
CREATED="2008-02-21T06:32:00">
            <METS:FLocat LOCTYPE="URL" xlink:type="simple" xlink:href="TS_cerp.xml"/>
          </METS:file>
          <METS:file ID="TS_TEST_2" MIMETYPE="application/vnd.ms-outlook" SEQ="1"
SIZE="228000"  CREATED="2007-11-14T06:32:00">
            <METS:FLocat LOCTYPE="URL" xlink:type="simple" xlink:href="TS_cerp.pst"/>
          </METS:file>
          <METS:file ID="TS_TEST_3" MIMETYPE="application/zip" SEQ="1" SIZE="336"
CREATED="2007-02-22T06:32:00">
            <METS:FLocat LOCTYPE="URL" xlink:type="simple"
xlink:href="TS_cerp_metadata_narrative.zip"/>
          </METS:file>
          <METS:file ID="TS_TEST_4" MIMETYPE="application/zip" SEQ="1" SIZE="1300"
CREATED="2008-02-22T06:32:00">
            <METS:FLocat LOCTYPE="URL" xlink:type="simple"
xlink:href="TS_subject_sender_log.zip"/>
          </METS:file>
          <METS:file ID="TS_TEST_5" MIMETYPE="application/zip" SEQ="1" SIZE="21"
CREATED="2008-02-19T06:32:00">
            <METS:FLocat LOCTYPE="URL" xlink:type="simple"
xlink:href="TS_cerp_EAD.zip"/>
          </METS:file>
            <METS:file ID="TS_TEST_6" MIMETYPE="application/zip" SEQ="1"
SIZE="190000"  CREATED="2008-02-22T06:32:00">
              <METS:FLocat LOCTYPE="URL" xlink:type="simple"
xlink:href="TS_cerp_Parser_directory_tree.zip"/>
            </METS:file>
          <METS:file ID="TS_TEST_7" MIMETYPE="application/xslt+xml" SEQ="1" SIZE="6"
CREATED="2008-02-22T06:32:00">
            <METS:FLocat LOCTYPE="URL" xlink:type="simple" xlink:href="TS.xslt"/>
          </METS:file>
          <METS:file ID="TS_TEST_8" MIMETYPE="application/xml" SEQ="1" SIZE="8"
CREATED="2008-02-22T06:32:00">
            <METS:FLocat LOCTYPE="URL" xlink:type="simple"
xlink:href="TS_mets.xml"/>
          </METS:file>
        </METS:fileGrp>
```

```
  </METS:fileSec>
  <METS:structMap ID="S1" LABEL="DSpace" TYPE="LOGICAL">
    <METS:div ID="div_1" DMDID="dmd_1" TYPE="DSpace Item" LABEL="DSpace">
      <METS:div ID="div_2" TYPE="DSpace Content Bitstream">
        <METS:fptr FILEID="TS_TEST_1"/>
        <METS:fptr FILEID="TS_TEST_2"/>
        <METS:fptr FILEID="TS_TEST_3"/>
        <METS:fptr FILEID="TS_TEST_4"/>
        <METS:fptr FILEID="TS_TEST_5"/>
        <METS:fptr FILEID="TS_TEST_6"/>
        <METS:fptr FILEID="TS_TEST_7"/>
        <METS:fptr FILEID="TS_TEST_8"/>
      </METS:div>
    </METS:div>
  </METS:structMap>
</METS:mets>
```