THE COLLABORATIVE ELECTRONIC RECORDS PROJECT SUMMARY

TABLE OF CONTENTS

Citation
Adgent, Nancy and Lynda Schmitz Fuhrig. (2009). "The Collaborative Electronic Records Project Summary." Sleepy Hollow, NY and Washington DC: The Collaborative Electronic Records Project.

In August 2005, the Rockefeller Archive Center (RAC) and the Smithsonian Institution Archives (SIA) launched the three-year Collaborative Electronic Records Project (CERP) to develop the methodology and technology for managing and preserving born-digital materials in archival collections. The project's primary objectives were to produce management guidelines and technical preservation capability that would enable archives and manuscript repositories to make electronic information accessible and usable for future researchers, and to share findings and products with depositors, peer institutions, and other interested non-profit groups. Differences between SIA and RAC contributed to the CERP's applicability to a wide range of institutions. SIA is both the institutional archives and the Smithsonian's records manager, serving all of the contributing units. It typically receives electronic records as part of mixed media transfers five or more years after the records have become inactive. SIA has accessioned born-digital material for almost fifteen years. On the other hand, RAC has no control over its depositors, does not own all the records it holds, and has not deliberately acquired born-digital records, although a few floppy disks and CDs have been accessioned along with paper documents.

Soon after embarking on Phase I, the focus narrowed to email for several reasons: 1) other projects were addressing website preservation; 2) CERP funding, staffing, and time limitations precluded a comprehensive approach; and 3) email preservation was urgently needed. The enormous quantity of email generated makes better management of digital communication economically advantageous. With the advent of email came a change in organizational roles. In most companies, every employee assigned an email account acts as a file clerk and records manager, and has the ability to create, destroy, mismanage, and improperly use email. Facing regulatory requirements and exposure to lawsuits, companies, particularly non-profits, cannot afford to ignore privacy, security, and rights ownership issues arising from email. Compiling, communicating, and enforcing an organizational email policy is essential to preserving a company's records for posterity and protecting its financial and intellectual assets.

Understanding donor institutions' electronic records organization and environment is important, so the project began with interviews of selected staff members at various depositing units. Based on the findings and research, CERP developed best practices guidance for creating, managing, transferring, and preserving electronic records; the documents may be downloaded from the project website at http://siarchives.si.edu/cerp. If donors adopt the guidelines, archives are more likely

to receive electronic records with authenticity and integrity intact.[1] Ideally, donor and depositor organizations should establish policies and procedures for generating and saving electronic information long before transferring records to an archive. Although some archives may not anticipate accessioning email for several years, planning for the transfer and preservation of born-digital records should begin as soon as possible.

While the CERP work was intended to apply only to records retained for historical research purposes, organizations may wish to consider using portions of guidance documents for other, current records. Determining which records to keep and for how long may vary from one organization to another. Whether a record is electronic or paper, its **content** determines its value and retention period. When a record is scheduled for destruction, it should be disposed of properly in order to ensure that it truly is no longer recoverable. CERP archivists developed records retention and disposition guidelines, also available on the project website.

During the project, many issues surfaced, including personal and confidential messages mixed with business email, missing attachments, lack of file order, deteriorating media, obsolete software, and unknown formats. In the course of testing small caches of email, the CERP team used a variety of freeware and commercial, off-the-shelf software to identify, assess, and convert formats. In the final phase, information technology consultants developed a parser to convert email to an XML preservation format, and customized an ingest module for depositing email and related metadata into the digital repository, DSpace. Serendipitously, CERP teamed with the North Carolina State Archives to refine and further test the preservation schema.

As the project concluded in late 2008, CERP had produced best practices guidelines, a workflow outline, evaluation of software tested, SIP/AIP/DIP models, a software tool that preserves email accounts together with their messages, and a customized DSpace ingest module, and had parsed more than 89,000 email messages with a success rate of 99 percent. The CERP website will remain viable indefinitely and occasionally updated; however, the research team has disbanded, thus troubleshooting and consultation cannot be provided.

[1] Authenticity means a record that is what it purports to be, i.e., includes the email with its attachments and transmission data, and that it was created by the credited author. Integrity is confirmation that a record has not been altered, intentionally or accidentally, since its creation or receipt.

The CERP Team

Rockefeller Archive Center	*Smithsonian Institution Archives*
Darwin Stapleton, Principal Investigator	Riccardo Ferrante, Principal Investigator
Nancy Adgent, Project Archivist	Lynda Schmitz Fuhrig, Project Archivist

Consultants Steve Burbeck & Lawry Persaud

COLLABORATIVE ELECTRONIC RECORDS PROJECT OVERVIEW

When the Collaborative Electronic Records Project (CERP) started, the archival community was only beginning to address electronic records issues, and few, if any, repositories were ready to tackle email archiving. Through this summary of CERP activities, the team is sharing "lessons learned" with a non-technical audience that may include archivists, records managers, records donors and depositors, and other interested non-profit institutions. For other CERP publications, updates, and additional information, see http://siarchives.si.edu/cerp. These products will remain available on the CERP website indefinitely for adoption and modification by any non-profit organization. The website will be updated when warranted.

Planning, Funding, and Staffing

The project 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 dearth of electronic records archiving theory and practice. The Rockefeller Foundation partially funded the CERP grant proposal, and the Rockefeller University (at the time the RAC's parent institution) committed additional resources. Nevertheless, the total was only approximately half the amount estimated for completion of the project as proposed, and plans to hire a senior systems engineer were dropped. During Phase II, CERP contracted with IT consultants Dr. Steve Burbeck and Lawry Persaud to perform some of the tasks originally planned for the systems engineer.

Because RAC did not have information technology staff as did SIA, project management was determined to be the responsibility of SIA's Information Technology Archivist/Electronic Records Program Director, Riccardo Ferrante. A Steering Committee was formed that included the two founders, RAC Assistant Director, and consultants Dr. Charles Dollar and Dr. Gregory Hunter, the latter two pioneers in the digital archiving field. After Dr. Hedlin retired, first the Acting SIA Director Tom Soapes, then the new Director, Anne Van Camp, replaced Hedlin on the Steering Committee, and later, Margaret Hedstrom, Associate Professor in the University of Michigan's School of Information, was added to the Committee.

In August 2005, each institution hired an archivist specifically for the project, and Stapleton, Dollar, and Ferrante publicized the project plans in a session at the Society of American Archivists annual meeting.

Phase I: Surveying the Situation

As both CERP archivists were new to their institutions, an initial orientation period was required to learn about current and potential donors and depositors and how the respective institutions, SIA and RAC, operate. With electronic records archiving still in its infancy, considerable time researching pertinent resources and reading applicable literature was necessary before launching the survey phase. Each CERP archivist devised a set of questions to guide the information-gathering process based on research into electronic records management issues and common sense thoughts about information archivists would need to transfer and process email; however, both RAC and SIA archivists refined and supplemented the questionnaire after early interviews. 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). In order to minimize the impact on depositor's time, RAC conducted only one visit to each participating depositor. SIA was able to make repeated visits to all contributors.

Major Findings

Interviews confirmed that a significant percentage of electronic records have already been lost through inadequate organizational procedures and absence of records retention policies in some cases as well as the lack of long-term technical preservation methods. Other results included:

- Much institutional history exists solely in electronic form and is not being systematically preserved
- No records manager or records management policy
- Email not recognized as an official record
- Lack of employee instruction in email creation, organization, and retention
- Paper file and folder naming standards not applied to electronic documents
- Personal messages mingled with business correspondence
- Some email systems used for desired email records are no longer in operation or are otherwise unavailable because the records pre-date

the organizations' current software and operating systems by several years
- Many attachments reside on a networked server rather than in the email system or on the email account owner's desktop hard drive and may no longer be accessible
- Email retention is dictated by IT storage capacity and backup policy

Results

From the surveys, CERP summarized the range of software applications in use and organizational practices for use in developing best practices guidelines and technical preservation solutions. After surveying RAC depositors, a comprehensive list of Rockefeller and related entities (including a summary of each organization's work and key personnel contact information) was compiled for future use in pro-actively soliciting electronic records while they are viable. "*Depositor Survey—Electronic Records Status*," used to determine depositors' electronic records environment and transfer readiness, is on the CERP website.

Developing Guidelines

Based on the conditions found during depositor interviews, RAC and SIA each developed best practices guidance to assist depositors and archivists with email management. Because RAC does not receive email directly from active email systems in contrast with SIA, which receives email from both obsolete and active email systems, procedures and guidelines were tailored for the different archiving environments. RAC's guidelines address issues and trends that corporate officers and managers of its depositing organizations need to consider, including records management principles, financial accountability, legal precedents, regulatory requirements, and operational needs and security. Legal cases are cited and examples of email management policies are listed. Most RAC depositors do not have a dedicated records manager, so to assist employees whose duties include that function, guidelines offer basic instructions about the records management role, how to determine which records are permanent, and how long to retain different categories of records.

SIA met with its Records Management Team to review the findings from the surveys and to select accounts for transfer for the project. SIA created transfer documentation that 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.

Many small archives and their donors have not trained employees in proper email creation, organization, and retention, thus a section in RAC's "*E-Mail Guidelines for Employees*" discusses etiquette and unacceptable use. A sample "*E-mail and Internet Policy Acknowledgement Form*," a glossary, and a list of resources are part of the guidance publication. SIA issued email guidance defining what makes email a record, tips for weeding email accounts, and some consequences of poor email management.

Results

The RAC's "*E-Mail Guidelines for Managers and Employees*" was published in a paper format, and is also available on a CD and as a download from the project and RAC websites. SIA published "*Responsible Recordkeeping: Email Records*" and "*Email Guidance*" documents for its depositing units and posted both to the CERP and SIA websites.

Phase II: Transferring, Testing, and Processing

Choosing Testbed Depositors, Accounts, and Transferring

Once CERP acquired enough information to assess the situation, RAC and SIA obtained cooperation from selected depositors in identifying and capturing email for use as testbed material. Although selecting, capturing, and some testing occurred during Phase I, the latter continued into Phase II and earlier guidelines were revised based on the knowledge gained while transferring, testing tools, and processing email accounts. Phases II and III became more technical with devising and refining functional and system requirements and developing products such as the preservation parser.

In accordance with its commitment to the two RAC email testbed depositors, all information that could identify the messages, creators, recipients, or offices of origin would remain confidential. RAC altered a standard Deed of Gift form into a Testbed Deposit Form, signed by both parties, to reflect the agreements. RAC also agreed that at the end of the project all testbed messages would be completely wiped from RAC computers and servers, copies on removable media would be properly destroyed, and the originals would be returned to the

depositors. All SIA testbed material was kept confidential during the pilot. Some material was accessioned by SIA at the conclusion of the project. The remaining material is being destroyed.

During the transfer process, CERP wanted to address issues involved with appraisal, accessioning, format identification and migration, media refreshing, and preservation, and doing so required considerable documentation. Some of the standard archival subjects CERP investigated were:

- Appraisal
- Authenticity
- Integrity
- Access
- Processing workflow and time

Appraisal for CERP testbed material varied by depositor. Before CERP started, one of the RAC testbed depositors had copied a former staff member's messages onto CDs. These dated back to 2001 and contained several email clients, some in multiple versions, including AppleMail, Eudora, GroupWise, Lotus Notes, Mozilla/Firefox, Outlook/Exchange, Pegasus Mail, and Simeon for MacPPC. Considering that the creator was a corporate officer and department head, RAC accepted the CDs without viewing them on the assumption that all the material would have historic value and merit permanent retention.

The second RAC depositor was in the midst of restructuring and was closing two grant-making units. There was discussion about the general contents of various Inbox folders with the two program officers involved, and together they opened a small percentage of messages on their desktops, and determined which folders contained the information RAC had kept in paper format for prior years. This depositor allowed its IT staff and the RAC CERP archivist to capture those pre-determined Outlook Inbox folders from their server onto CDs in PST format.[2]

A third RAC testbed consisted of twenty-nine previously donated CDs containing 18,000 scanned files that had become difficult to access because the software program used for the scanning project in the 1990s is no longer supported by the vendor. Paper originals had been destroyed. Subsequent research found that of the four archives holding a copy of the data, only the

[2] 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. A PST file 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.

RAC's was viable, thus its appraisal decision to attempt preservation was intuitive.

The three SIA testbeds consisted primarily of Outlook Exchange email accounts and were from administrative, financial, and scientific research units. At the beginning of Phase II, 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.

Other SIA transfers were conducted by an Office of the Chief Information Officer (OCIO) staffer and the CERP project manager using Microsoft Exmerge for Outlook for secure transfer.[4] 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. 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

[3] Smithsonian Institution follows strict computer technology protocols as defined in various federal guidelines and best practices.
[4] Exmerge is a MS Exchange utility program that can extract data from mailboxes on an Exchange Server.

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 accession 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.

Testing and Processing

While testing CERP intended to answer processing questions, some routine and others peculiar to electronic records, including:

- Should CERP impose an order on a collection that was not organized by the depositor?
- How will electronic records be correlated in accessioning documentation and finding aids with paper and other analog records from the depositor?
- Will the archive commit resources to redact personal, sensitive, confidential, SPAM, and duplicate messages?
- How should CERP isolate or remove viruses?
- How will attachments be linked to email messages throughout processing?
- How does CERP determine if and when native attachments should be migrated to new and/or stable formats?
- How does CERP determine when removable media should be refreshed?

RAC and SIA's first processing step was to make two copies of the original, native email being transferred, resulting in three sets – the original, a redundant copy, and a working copy. All processing was performed on the working copy. Then RAC compared the folder and file size of the original to the copies, and opened each to sample a percentage of the messages for complete and accurate transfer. RAC first viewed some testbed folders in Notepad, but found it slow to display even on a small batch less than 7 MB, and it would not open batches over 100 MB. Viewing in Internet Explorer was faster, although still slow on larger batches.

Next RAC and SIA conducted virus scans with the commercial, off-the-shelf anti-virus software used by the respective institutions for non-CERP work. Results of

authenticity and integrity verification and virus checks were recorded on the Electronic Records Verification Form at RAC while SIA relied on a metadata narrative that indicated 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 by documenting tools used and conversion procedures taken.

Virus programs differed in their findings and, in one case, the viruses found in emails or their attachments could not be quarantined or cleaned when moved from CD to a PC desktop at RAC. Luckily, the viruses detected were old (RAC was testing email created 2 to 6 years prior) and posed no threat to current operating systems. SIA chose not to process the email messages flagged with viruses and documented this. Ideally, email should be cleaned by the depositor before transferring; however, transferred email should be scanned for viruses and placed on a secure, non-networked desktop or server rather than on ones used for regular daily work.

Because RAC's transfers were from external organizations, it developed the Electronic Records Transfer Form on which the archivist could document the collection, record group, and series names, accession number, Archival Information Package (AIP) number, name and title of the email creator, date range of the batch being transferred, type of content (email, spreadsheets, database, etc.), format (Outlook), type media (CD, server, etc.), source (desktop, server, portable device), and the destruction date.[5] RAC also modified an Accession Form for email. As with all forms developed during the project, both are meant to be maintained electronically, and they are included in guidance documents available on the CERP website. SIA used its own metadata template referenced above for this documentation.

Another decision facing archives is whether to accept email from depositors who have not deleted personal, sensitive, confidential, and SPAM messages. If unsorted email is accessioned, the archive then has to determine whether to use an archivist's time for this purpose. RAC compiled a list of approximately 50 words that could identify a message as non-business and a separate list for business-related terms. On a batch of 5,170 messages, using the lists identified an average of 224 messages per hour whereas manually reading the subject lines (and opening messages in question) produced 247 per hour. Neither institution

[5] The SIP/AIP/DIP concept we used is based on the OAIS Reference Model adopted by the International Standard Organization as the standard for long-term preservation and access of digital materials in a repository. See http://public.ccsds.org/publications/archive/650x0b1.pdf for more information.

attempted to delete duplicates; this action may be feasible in the future if an automated tool is available.

At SIA, 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.

Attachments pose still another preservation challenge. Because of the variety of attachment file formats, the attachments may obsolesce at a rate different from the email messages. The question facing archives is whether the organizations undertake more time-consuming work to assess the attachments' long-term format viability, and potentially extract and migrate them to stable, recommended preservation formats. SIA did this by extracting the attachments using Aid4Mail, 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.[6]

Once the attachments were extracted, the file format identification tools JHOVE and DROID were applied to the collections.[7] 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.[8] 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.

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

[6] See http://www.siarchives.si.edu/cerp/RAC_SIA_CERP_tools_V2.pdf for an evaluation of tools used by CERP.
[7] 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.
[8] Troubleshooting was required due to java and configuration file issues on the SIA workstation. The Harvard team was very helpful.

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?[9] 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.

RAC chose to simply identify and convert formats without determining obsolescence, reasoning that the migration would be necessary at some point and the conversion would likely be less problematic if done sooner instead of later, particularly since RAC expects to accession most email collections more than 3-5 years after creation. Both SIA and RAC kept the original attachments with the source email.

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. [10] ArcGIS was installed on the PC in 2006, and

[9] Base64 is a binary-to-text encoding schema. Others include hexadecimal, quoted-printable, and BinHex.
[10] PCs running Windows/Microsoft Office allow for the automatic substitution of another font within a document for the one not installed on the machine.

the archivist was not aware of the font replacement change. There was an attachment within the attachment too.

One standard archival decision – whether to impose an order on a collection when no original organization was done – needs to be addressed for email accessions also. This may be the case when the transferred email does not arrive in the context of an email account or the account had only an Inbox folder and no further structure. Depending on the size of the email account being transferred, organizing email will likely be too time-consuming for archivists. Imposing organization on the messages also raises the issue of original order. Each archive will need to make a decision regarding the procedure it wants to follow, and the procedure may vary according to the importance of a collection. SIA kept the structure that was used by the account holder.

Next RAC and SIA tackled the task of converting email to XML (eXtensible Markup Language), CERP's selected preservation format. 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.[11] Other institutions such as the Antwerp City Archives and National Archive of the Netherlands used XML as well for their email projects. 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[12]

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.

CERP's first IT consultant was hired during Phase II, and after the decision was made that the parser prototype under development would require MBOX format for the incoming messages, CERP investigated software that would convert the original, native email formats into MBOX.[13] Because RAC's testbed email

[11] XML is currently used in many websites and its web page display is achieved through XSLT, CSS, and/or Javascript.
[12] Available online at http://sourceforge.net/tracker/index.php?func=detail&aid=1946019&group_id=85722&atid=577089. Accessed Dec 1, 2008.
[13] MBOX is a generic format likely to be viable for decades. . 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

consisted of a large variety of email applications, Aid4Mail worked better than other tools to convert the native, source messages (other than Outlook PSTs) from MBOX format into EML format for processing, then back to MBOX for conversion into XML preservation format.

The conversion from MBOX to EML was done because the MBOX display is too difficult to use for sorting personal, confidential, and sensitive material, and is not an efficient use of an archivist's time. For processing Outlook email, both RAC and SIA used MessageSave to convert the proprietary PST format into MBOX. Due to the processing time involved and the possibility of mistakenly overlooking recordworthy material, SIA did not sort out messages. RAC, on the other hand, produced a "researcher use copy" which has had the personal, sensitive, confidential, and junk mail removed.

Both RAC and the SIA produced finding aids for testbed material, and the SIA created the online finding aids as EAD using NoteTabPro initially, later using o**X**ygen.

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.* Development of processing workflow and tools continued into Phase III.

Results

Throughout the testing phase, CERP documented changes when particular software was used, what actions were taken, and any anomalies. From this, CERP continued drafting transfer guidelines and the RAC developed Records Retention and Disposition Guidelines, Electronic Records Accession, Migration Schedule, Transfer Guidance and Documentation, and Verification Forms. CERP successfully processed email using standard archival concepts of appraisal, accessioning, original order, organization, description, and conservation assessment.

---

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. For more information, see http://en.wikipedia.org/wiki/Mbox.

Phase III: Preserving, Storing, and Retrieving

Preserving

Early in the project, CERP decided it would pursue email archiving as accounts rather than as individual messages, chiefly because: 1) the sheer volume precludes using scarce archival resources to preserve each message and document its contextual relationships; and 2) the value of preserving email messages "in situ" resolved issues of original order and overall metadata and documentation. Dr. Steve Burbeck, IT consultant, began developing a parser to translate email from the generic format (MBOX) into XML for long-term preservation. Serendipitously, he began talking with David Minor of the North Carolina Department of Cultural Resources who had designed an XML schema (http://www.archives.ncdcr.gov/mail-account) for use with that state's email account preservation project, EMCAP.[14] That project, funded by the National Historical Publications and Records Commission (NHPRC), and involving the state governments of North Carolina, Kentucky, and Pennsylvania, is also specifically addressing email preservation.

Collaborating on the schema and testing it on both projects' testbed email accounts proved very beneficial for CERP and EMCAP as differences among the accounts presented a wide range of challenges to develop a parser that would work for the vast majority of situations. The parser converts email messages, associated metadata, and attachments from MBOX into a single preservation XML file that includes the email account's organizational structure. The parser was successfully used on both Windows and Linux operating systems. A web-based user-friendly interface was used on Windows. The parser outputs the parsed email in one XML file, each attachment over 25 KB into a separate XML file, each bad message into a separate file, and a comma separated values Message Summary file, also known as the Subject-Sender log. The Message Summary file includes basic metadata about the Bad Messages in each batch processed such as To, From, Date, Subject, the unique message identification number assigned by the parser, a hash code used to ensure authenticity, and the first error in each bad message listed in the Summary. The Subject-Sender log presents the same information for all messages in the account processed.

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

[14] See http://www.w3schools.com/Schema/default.asp for more on schema.

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.[15] 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. After six months of code changes and tweaks, the parser was installed at SIA and later at RAC. 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.

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.

Storage

CERP decided to use DSpace as the testbed digital repository, primarily due to its large user community, maturity, and its use already at the Smithsonian Institution Libraries and Rockefeller University (at that time, the RAC's parent organization).[16] CERP used a technology adviser with expertise in DSpace.

[15] 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.

[16] DSpace is Open-source content management software developed by MIT and Hewlett-Packard for use in preserving, storing, and allowing access to digital information. Its community of users, primarily academic institutions, determines their own policies for deposit, storage, and retrieval. See http://www.dspace.org/.

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.[17] 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.*

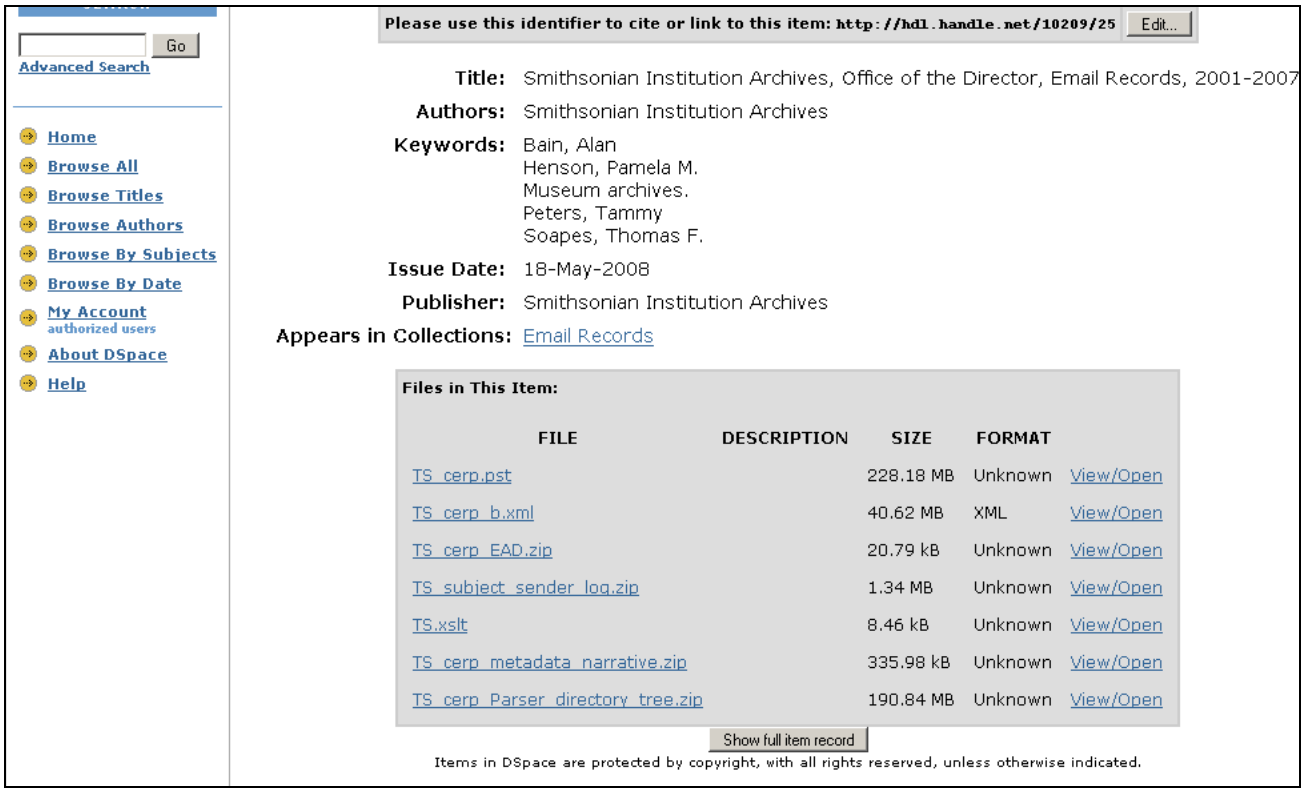


Figure 1

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.

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

The METS file was structured in the following way:

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 document), 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>[18]
<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:identifier.other>Accession 07-109</dc:identifier.other>

<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>

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

File Section
For each file described by this METS file, by FileGroup <FileGrp>.
ID
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 AIP stored in DSpace consists of:

- source format email account (.pst, .msg, etc.)
- MBOX format email account (preliminary preservation transformation)
- Any other format conversion such as EML
- Preserved format account (XML)
- Metadata – administrative with preservation assessment & descriptive including narrative Finding Aid and attachment format reports
- Parser output – Directory Tree.zip, Bad Messages, Subject-sender log.zip
- METS.xml used for ingesting the AIP into DSpace
- File Name (e.g. John Doe E-Mail) METS.xml (administrative & descriptive metadata encoded in METS)
- XML stylesheet, used to facilitate later display of the preserved account

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.

RAC stored original and preservation copy CDs in Tyvek envelopes within archival CD boxes housed in a secure vault with temperature and humidity controls set at 50 degrees Fahrenheit and 40% relative humidity. For non-testbed materials, the recommendation is to store another, redundant copy offsite in a proper physical environment. SIA retained their original and preserved accounts online with redundant copies on an external drive and tape. CDs also were made.

Retrieving

In addition to searching Dublin Core elements, 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.

Some of the questions to be considered when developing retrieval guidelines and permissions include:

- Who has permission for what tasks – access, modification, viewing?
- Who has access to which components of the DIP, e.g. will all archivists be allowed to access the native, source email?
- How will researchers view files – on a dedicated, non-networked desktop, on a secure server, etc.?
- How will the collection be protected from malware, viruses, piracy, misuse, etc.?
- Will researchers be allowed to print, copy, save, or email archived messages?
- In what ways will depositors' access rights differ from researchers' rights?
- Do you want users to be able to search for keywords in individual messages or browse messages within a particular series, folder, etc.?

- Will search terms be based on standards such as Library of Congress Subject Headings (LCSH)?

Results

CERP achieved a 99+ percent success rate in parsing RAC and SIA messages. More than 36,000 SIA Outlook messages totaling approximately 2.7 GB and more than 46,000 older, eclectic RAC messages totaling approximately 500 MB were parsed. The SIA sets contained more attachments than the RAC sets, thus the discrepancy in size. Parse rates equaled a rate of about one-fourth GB per hour on an IBM laptop. Parsing time varies depending on the processing power of the PC, the messages' attachment content and size as well as the "legality" of the email messages themselves.[19] CERP deposited testbed email accounts into DSpace using the CERP METS ingest module and was able to retrieve them using the ten key elements on the METS form. The parser and a Parser Installation and User Guide will be posted on the CERP website.

A Wish List

As a rule, grant-funded projects rarely have time or funds to refine their deliverables, and CERP is no different. CERP would like its work carried forward in several areas:

1. Migrate the Parser from SmallTalk to a different technology platform to make it more easily used by non-technical staff
2. Automate EAD finding aid creation (this may be unnecessary as more archives gravitate toward Archivists Toolkit and similar collection management software applications that have the capability of converting to EAD).
3. Automate METS file generation
4. Automate AIP assembly
5. Enhance METS Import Utility to provide full text indexing
6. Searching: Select accounts based on search criteria match to emails within multiple accounts.
7. Retrieval: Display emails that meet search criteria individually rather than forcing the researcher to browse through the account from beginning to end

[19] An email is considered "legal" if it meets the RFC 2822 syntax standard established by The Internet Society for email messages. For example, dates must appear in day/month/year sequence.

8. Retention/Destruction: Automate a calendar-based notification system that would alert an archivist when particular files need to be migrated, refreshed, or destroyed.

Lessons Learned

Planning, Funding, and Staffing

Ideally, an electronic records archiving team would include at least one person with traditional archival skills and knowledge and one information technology staffer, and both would know enough about each other's field to discuss methods and issues and understand current literature about the topic. RAC's lack of IT staff during the project slowed progress and contributed to inadequate computer systems infrastructure. On the other hand, not having funds to hire a CERP systems engineer forced CERP to locate IT consultants who could accomplish the tasks of developing a parser and customizing DSpace ingest. CERP was fortunate to find two very capable consultants who achieved the goals, and very likely did so better and quicker than one all-purpose engineer would have.

Analytical and organizational skills are important for the archivists; basic knowledge of HTML and XML and standards such as EAD, DACS, TRAC, and OAIS is very helpful, as are the patience and willingness to experiment with software. Because some depositors will likely need to be educated in managing email, archivists should have the ability to prepare presentations and communicate with groups of depositor employees as well as to develop training materials and instruct selected individual employees on both the depositor's and the archive's staffs.

Surveying

- Claiming even an hour of time from busy colleagues, not to mention unaffiliated depositors, requires much persistence, patience, and flexibility from archivists.
- There is NO substitute for in-person interviews. Merely requesting that depositors complete a survey form will not elicit the responses required for a thorough analysis of a depositor's electronic records situation. During every visit, both the SIA and RAC archivists learned details simply by following up on conversational twists and by seeing the Inbox organization and the email management process actually in use.
- With the rapid changes in technology, the survey on the CERP website will need to be updated.

Developing Guidelines

- Beginning with imperfect guidelines now is better than procrastinating.
- The deed of gift/deposit should include a clause absolving the archive of liability in the event a depositor's employee's personal email is inadvertently captured and seen by a researcher.

Transferring, Testing, and Processing

- Documentation, documentation, documentation in detail is very important.
- It is better to use at least two virus checking applications, and experimenting with several applications may be necessary to find one that works most accurately for email from particular depositors.
- Progress is hampered if using outdated computers and ones with inadequate RAM.
- Testing as wide a range of email applications and creation dates as is practical will improve the processing success rate.
- The existence, location, and formats of electronic records on deposit, or that are not to be retained permanently for any reason, will need to be carefully documented so that all versions and copies on hard drives, servers, and removable media can be completely sanitized or properly destroyed when the retention period ends. A proper certificate of destruction will need to be completed, signed, and given to the depositor.
- Opening links referenced in email and migrating them to a preservation format proved too time-consuming to be feasible except possibly on a very valuable collection. Researchers may find at least partial information on the Internet Archive/Way Back Machine website. This situation correlates to paper correspondence in which the writer references an event or source not explained within the collection.
- Sorting out non-business messages and deleting duplicates is too time-consuming to be done on the large volumes of email expected – unless automated tools are developed.

Preserving, Storing, and Retrieving

- No parser will work on 100% of the email ever created.
- Various XML software programs experienced difficulty in opening large XML files for validation.
- Archivists will have to manually address a small percentage of problem messages.
- Talking with people working on other electronic records projects is beneficial even if no collaboration develops as both teams learn.

- Determining search criteria and metadata tags will vary greatly by organization, by archive, and by collection.
- How future researchers will construct queries or use search features is speculative, and we can expect the unexpected, meaning that what worked for this project will not suit all repositories and all researchers.

Sharing Experiences

- CERP's experience does not provide answers to every problem.
- Archivists, donors, and publishers of archival literature are thirsting for information about email archiving.

What Can an Archivist Do Now?

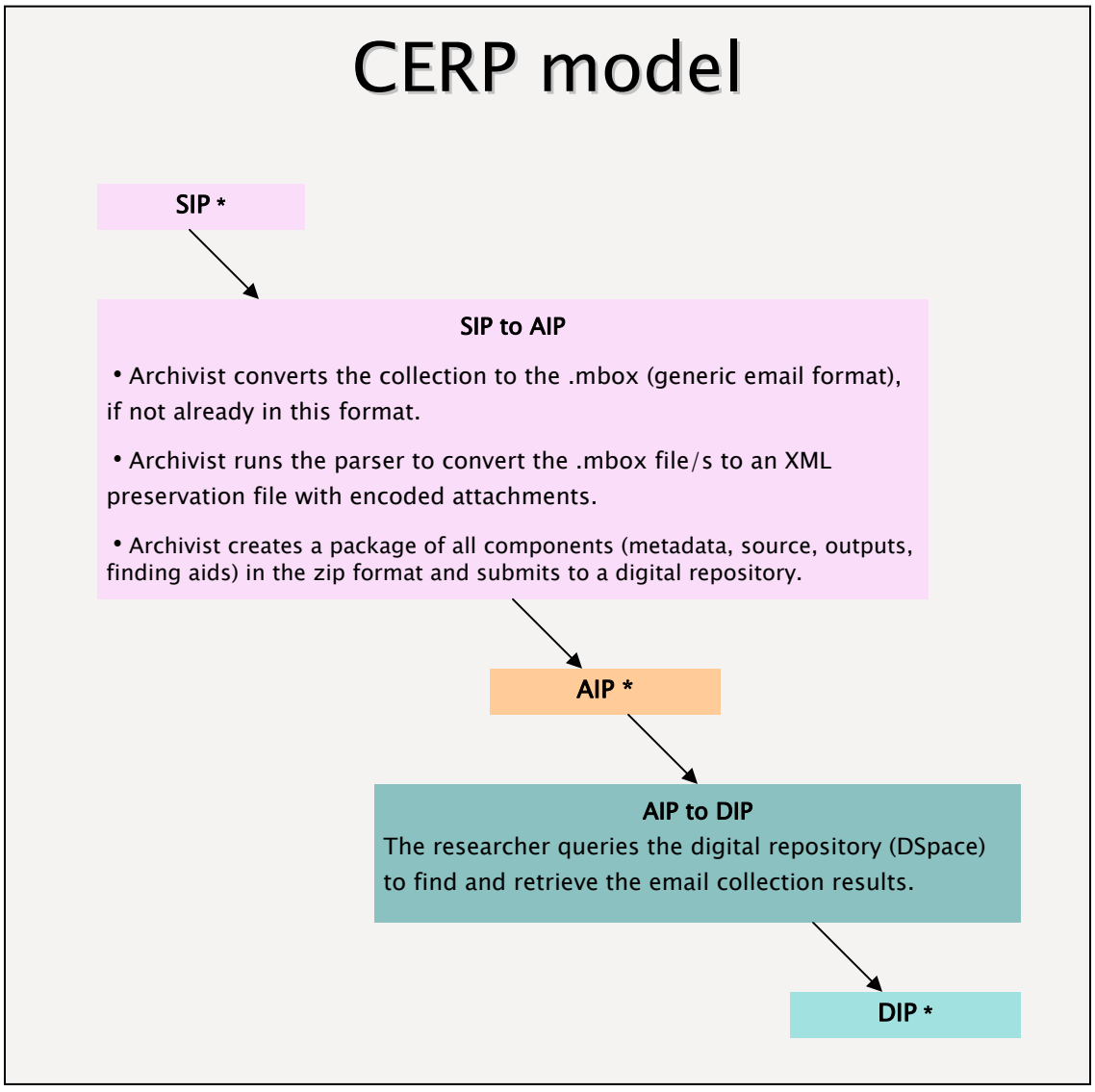
Many people, usually those in small institutions with no Records Management department and no electronic document management software system, have asked what they should do now – until their organization's management formulates policy, workflow, and budgets for proper email archiving. Here are a few suggestions:

1. Review the guidance documents on the CERP website and adopt the portions within your authority, expertise, and wherewithal to accomplish.
2. Discuss copying capability and storage capacity with IT staff. Perhaps they can copy Inboxes of people who create "records" at specified intervals (e.g. six months after fiscal year end or just before a person leaves), and save them on CD/DVD, external hard drive, dedicated server not used for other purposes, etc.
3. Organize your own Inbox into appropriate folders reflecting the file categories used for paper records and share your email management practices with colleagues informally. Often co-workers adopt desirable work habits by imitating their peers.
4. Do not keep personal messages in the same Inbox folder with business correspondence.
5. If your IT department instructs you to delete all email older than a certain date or to reduce the size of your Inbox, first copy folders containing official records to CD/DVD.
6. Persist in bringing to management's attention the need to establish organization-wide policies for email creation, organization, and

storage and to collect and preserve born-digital records as soon as they are no longer needed for daily work.

7. Review transfer guidelines on the CERP website and follow them to the extent possible.
8. Store electronic records on removable media in the proper housing and physical environment. See "*Care and Handling of CDs and DVDs – A Guide for Librarians and Archivists*" by Fred Byers, NIST Special Publication 500-252, Oct. 2003 http://www.itl.nist.gov/div895/carefordisc/.

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 E-mail	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 e-mail
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

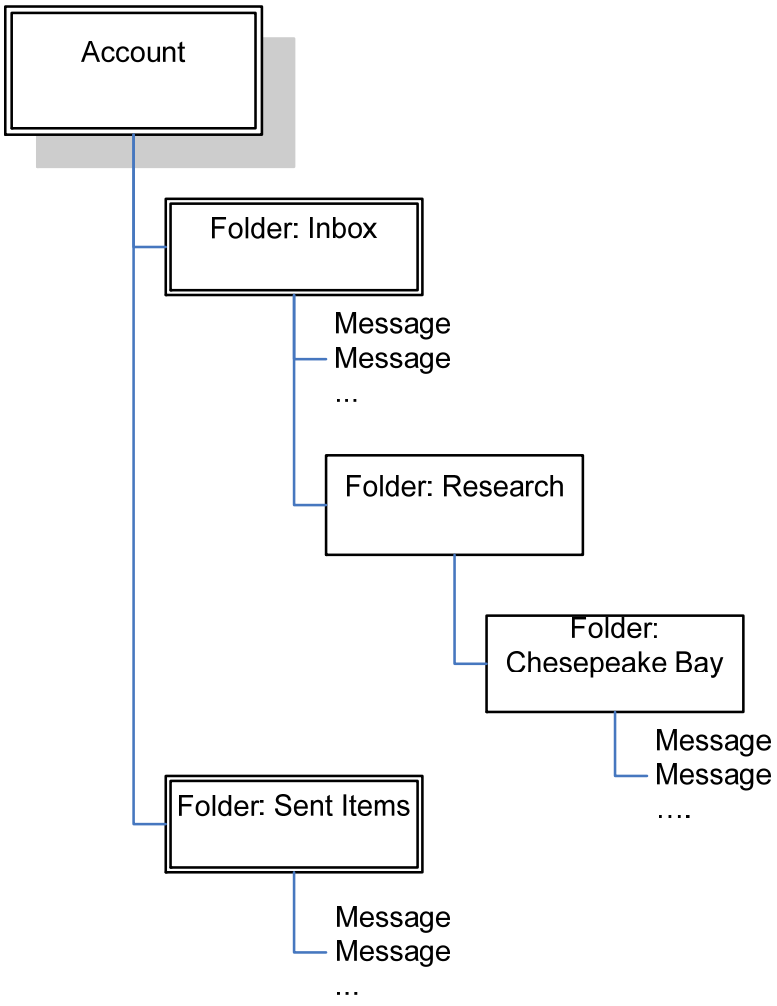
Metadata – Initial
- Metadata – administrative with preservation assessment
- Metadata – descriptive including narrative Finding Aid

AIP Assembly
- Source e-mail (.pst, .msg, etc.)
- Preservation e-mail
  - 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>
```

GLOSSARY

Active record:
A record in current use frequently in conduct of daily business.

Administrative Metadata:
Information needed to manage digital content and that is not part of the digital resource itself. Examples include acquisition date, copyright ownership, and disposition date.

AIP (Archival Information Package):
Originally accessioned digital content plus content converted to preservation format (such as XML) and associated metadata required for storage in a repository such as DSpace.

Archival record:
Information with legal, financial, administrative, or research value that should be kept permanently according to an organization's Records Retention & Disposition Schedule.

ASCII:
A text file where each character or space is represented by one byte encoded according to the ASCII (American Standard Code for Information Interchange) code. It preserves Latin-based alphabetical characters, punctuation marks, and some symbols and formatting.

ASP (Active Server Page):
This web page format uses scripting, normally VBScript or JavaScript code in combination with HTML, to dynamically generate a complete HTML page for display on the requesting web browser. The complete HTML is not generated until that page is requested by a web browser.

Audit Trail:
A record of actions performed on a computer system. It includes user identification as well as time and date information.

Authenticity:
A record that is what it purports to be and has not changed since its creation. Authentic e-mail includes the e-mail message as well as any attachments and its transmission data.

Born-digital:
Material (text, images, audio, video) that was created in a digital format. Not to be confused with digitized materials that have been converted from paper or other original type to a digital format by scanning or other methods.

CFM:
Cold Fusion template/page. Cold Fusion is a Macromedia web development application used to create dynamic web pages.

Convenience copy:
A copy of a record kept for reference and quick access.

CSV:
Comma Separated Values. Another name for comma-delimited text format. CSV preserves the data input (not formulae or formatting), allowing a spreadsheet or database to be recreated later.

DBF:
Database format used by various applications.

Descriptive metadata:
Information within and external to an electronic record that references selected components of its content for use in identifying or locating the record, such as a finding aid, a search term, or type media.

Digital Curation:
Management and preservation of digital objects (data generated in binary code) over their lifecycle of current and future use, ensuring the data retains its authenticity, access, reproducibility, and longevity. It includes selection, appraisal, intellectual control, redundant storage, data migrations, bitstream preservation, and metadata capture and creation.[i]

Digital obsolescence:
Digital data that was created in out-dated programs or operating systems or on old media that is difficult or impossible to access in the current digital environment.

Digital record:
Iinformation created or stored in a format that provides evidence of activities, events, decisions, programs, policies, or transactions. It may be born-digital or digitized.

DIP (Dissemination Information Package):
An information package, such as an e-mail accession or a journal, delivered from a digital repository upon request from an archivist or researcher.

Discovery:
Legal process in which one party to a lawsuit is required to furnish documents requested by the opposing side.

Disposition:
Routine, planned disposal of records by scheduled transfer (for permanent) or destruction (for non-permanent).

Document management system:
Computer software that files, routes, and retrieves documents created electronically regardless of the document's original format (Word, Excel, etc.).

DPI:
Dots per inch. A means of expressing the amount of information recorded in a digital image correlating to the resolution quality or density of the image.

DSpace:
Open-source content management software originally called Durable Space and developed by MIT and Hewlett-Packard for use in preserving, storing, and allowing access to digital information. Its community of users, primarily academic institutions, determines their own policies for deposit, storage, and retrieval; however, preservation is at the bitstream level with only a few formats renderable. See http://www.dspace.org/.

Dublin Core:
ISO/ANSI standard (15836/Z39.85) that defines metadata elements used to describe and provide access to online resources. Elements include title, creator, subject, publisher, date, etc. See http://dublincore.org/.

EAD (Encoded Archival Description):
EAD is the non-proprietary standard for encoding finding aids for use in an online environment.

ECM (Enterprise Content Management):
Use of technology to manage an organization's information flow from creation through storage. The term typically is used when referring to a company that provides software that captures, preserves, and retrieves electronic records. ECM also often includes management of digital rights, web content, and records retention.

E -Discovery:
Legal process in which one party to a lawsuit, or an organization subject to governmental regulation, is required to furnish documents generated and/or maintained in electronic formats to the opposing side upon their request.

8.3:
The MS-DOS file-naming convention of eight characters followed by a period (.) and three final characters. The three final characters are popularly used as acronyms for the file format of the electronic document. For example, "demo.ppt" is a Microsoft PowerPoint document. PPT would be the ".3" expression, or the acronym for a PowerPoint file.

ECPA (Electronic Communications Privacy Act)
Federal law that defines invasion of privacy regarding electronic communication, including e-mail, cellular telephones, pagers, etc.

Electronic document management system:
Computer program that enables an organization to manage its electronic documents from creation through storage and retrieval. [Note: this is not the same as archiving electronic documents.]

Electronic record:
Information created or stored in an electronic form that provides evidence of activities, events, decisions, programs, policies, or transactions. Electronic records include born-digital, digitized, and non-digital content such as video tapes.

Electronic signature:
According to the New York Electronic Signatures and Records Act, an electronic signature is "an electronic identifier, including without limitation a digital signature, which is unique to the person using it, capable of verification, under the sole control of the person using it, attached to or associated with data in such a manner that authenticates the attachment of the signature to particular data and the integrity of the data transmitted, and intended by the party using it to have the same force and effect as the use of a signature affixed by a hand."

EML:
E-mail format used by Microsoft Outlook Express and other e-mail applications.

Emulation:
Way of mimicking hardware or software so other processes think that the original equipment or system is still operating in its original form.

Encryption:
Method of hiding electronic information by encoding it so that only authorized persons who have the decryption code may access the data.

E-Sign (Electronic Signatures in Global and National Commerce Act):
Federal law that gives electronic signatures the same legal status as handwritten signatures with regard to electronic transactions.

Format:
Type of computer file, e.g. Microsoft Excel or JPEG image.

HTML:
HyperText Markup Language is a markup language for Web pages.

IT (Information Technology):
The system that handles information generated or stored through computers and telecommunications. Also known as Information Services (IS) or Management Information Services (MIS).

Integrity:
Confirmation that a record has not been altered, intentionally or accidentally, since its creation or receipt.

Internet Header:
Metadata viewable through e-mail software tools that gives information in addition to that shown in an e-mail message. The Internet Header gives IP addresses of sending and receiving computers, date and time stamps, and other details which may authenticate the message.

JPEG/JPG:
JPEG is a lossy compression technique for color images developed by the Joint Photographic Experts Group. File sizes can be reduced, but with a loss in detail. JPG is an alternate representation of JPEG.

LAN (Local Area Network):
A network of personal computers, usually within each location of an organization, that allows transmission of data within the network.

Life Cycle Management:
Retaining or destroying documents when they reach a pre-determined age and in accordance with government regulations, legal or financial guidelines, or an organization's internal policies regarding records retention.

MARC:
MAchine Readable Cataloging, a format for structured descriptive bibliographic, authority, classification, and holdings data. (Based on ANSI Information Interchange Format standard Z39.2). See http://lcweb.loc.gov/marc/.

MDB:
Format for Microsoft Access database (2003 and earlier).

MBOX:
A generic format for e-mail messages. All messages in an MBOX mailbox are concatenated and stored as plain text in a single file.

Metadata:
Internal metadata is information inherent within a digital document automatically produced when an electronic document is created, sent, modified, or received that describes its subject, date created, sender, recipients, etc. External metadata refers to preservation, technical, and descriptive information not part of the document itself that is created by a document creator, archivist, or other user. Metadata is used to identify, manage, preserve, and access digital information and includes format, size, accession source and date, disposal date, migration requirements, etc.

METS (Metadata encoding and transmission standard):
An XML format used for depositing text and image digital content and encoding its descriptive, administrative, and structural metadata necessary for managing digital accessions in a digital repository and for sharing that content with other repositories and users. A METS document is usually a required component of SIPs, AIPs, and DIPs.

Migration:
The process of transferring data from one electronic format to another, usually from older technology to newer. This is done to preserve information that might otherwise be lost as the old format becomes obsolete.

MIME (Multipurpose Internet Mail Extension):
The standard encoding method for e-mail attachments most frequently used.

.msg:
A proprietary binary e-mail format used by Microsoft Outlook.

Near-line storage:
Storing information in an electronic format apart from the e-mail system, such as on a desktop computer's hard drive or a shared drive. E-mail remains somewhat functional.

Official copy (also known as record copy):
Original record or a copy that is retained in compliance with an organization's Records Management Policy and Records Retention Schedule. If the e-mail is created within the organization, the sender usually maintains the official copy. When it is received from outside the organization, the primary recipient usually holds the official record.

Official record:
Information created or received in the course of conducting an organization's business, and required by law or deemed appropriate to be preserved, either short or long term.

Off-line storage:
Storing information outside an electronic environment, such as on paper copies, magnetic tape, optical disk, or computer-output-to-microfilm.

On-line storage:
Storage of e-mail, metadata, and attachments within the e-mail system currently being used by an organization. E-mail remains fully functional, i.e., it can still be forwarded, replied to, etc.

OAI (Open Archives Initiative):
An organization that developed and published application-independent interoperability standards to facilitate management and sharing of online content from harvested metadata. See http://www.openarchives.org/.

OAIS (Open Archival Information System) reference model:
Model serves as a reference for long-term preservation and access of digital materials in a repository: how digital objects can be prepared, placed in an archive, and stored, maintained, and retrieved. Many in the cultural heritage field have adopted it for their digital preservation efforts because of its flexibility and acceptance.

Parser:
A computer program that interprets digital data input such as e-mail text and converts it to XML or other format.

PDF (Portable Document Format):
Software developed by Adobe Systems that operates on several platforms (Mac, Windows, UNIX, etc.) and converts a variety of formats including Microsoft Word, Publisher, and PowerPoint, into a file that usually looks almost exactly like the original. The PDF version loses some automatically generated metadata and may lose some special formatting such as underlining. PDF is an open standard under the International Organization for Standardization (ISO) 32000.

Preservation Metadata:
Technical information required for managing and preserving digital assets over time to ensure the digital objects remain viable. It includes documentation of preservation actions such as migration, as well as collection and rights management information.

PST (Personal Storage File):
Microsoft Outlook proprietary format that creates one file containing all selected e-mail messages and attachments. It is stored outside of the e-mail server.

Record:
Formal or informal information generated within an organization or received by it during its course of business. A record may be in various forms whether printed or electronic, including book, CD/DVD, e-mail, instant message, map, memory card or stick, handwritten notes, memos, and sketches, photograph or other image, spreadsheets, audio or video tape, voice mail. (See Official Record.)

Records Management Policy:
A formal, written document containing an organization's procedures for managing records of its activities. It typically includes guidelines regarding which records to retain, the length of time they should be kept, the manner in which they should be organized, and the procedures for disposing of them or transferring them to an archive.

Records retention schedule:
A list of an organization's records by record type that indicates how long each type should be retained.

Refreshing:
The process of transferring data from one electronic media to another, usually from older technology to newer. This is done to preserve information that might otherwise be lost as the old media deteriorates or becomes obsolete.

Retention period:
An organization's pre-determined 'expiration' dates – the point in time when a record may be destroyed. Financial, legal, and governmental requirements, in addition to the organization's administrative needs and the historical value of the records, are considerations in establishing retention periods.

RGB:
Red, Green, Blue components of a color TIFF image

RMA (Records Management Application) or RMS (Records Management System):
Electronic document management system with an added feature that applies the organization's retention schedule to determine how long to retain a particular record. The purchasing organization usually works with the software provider to assign recognition identifiers (such as keywords in e-mail subject headings) and retention criteria.

RTF:
Rich Text Format. A format standard which embeds basic formatting instructions in an essentially ASCII document. Margins, font style, indentation and other formatting instructions are supported.

Schema:
An expression of data structure and content in tagged format, usually in XML, that enables machines to perform tasks ordered by human computer operators.[ii]

Security log:
A record of access, attempted access, and use of a computer system automatically kept by security software such as a virus protection program.

Signature line:
Lines of user-determined text, usually containing name, title, organization name, and contact details, set to be automatically entered by an e-mail client at the end of an outgoing message. [Note: this is not the same as an electronic or digital signature.]

SMTP (Simple Mail Transport Protocol):
Commonly used rules for e-mail transmission through the Internet.

Source file:
Digital files as originally created or deposited/donated to an archive or other repository.

Spoliation:
Unauthorized, whether accidental or deliberate, destruction of records pertinent to lawsuits or regulatory body investigations, or potential suits or investigations.

Structural Metadata:
Information about the divisions, views, extent, sequence, use, and relationship between parts of a compound object, such as pages and chapters of a book, table of contents, PDF file for download and printing, TIFF file for display, etc.

SIP (Submission Information Package):
Source data and relevant metadata provided to an archive by the data creator or a person or entity acting on the creator's behalf.

SWF:
Shockwave file format commonly referred to as Flash component, used by Macromedia's Flash player application. A popular plug-in, or supplemental application, used with web-browsers.

Tags:
Symbols used in electronic documents that instruct a program how to display the documents, e.g., font type and size.

TCP (Transport Control Protocol):
The rules that enable computers to communicate with each other through the Internet.

Text file:
An electronic file that can be read by many computer programs because it consists solely of ASCII characters and formatting.

TIFF (Tagged Image File Format):
A popular format for storing bit-mapped images; supports black-and-white, grayscale, and color images.

Unicode:
A character encoding standard developed by the Unicode Consortium. By using more than one byte to represent each character, Unicode enables almost all of the written languages in the world to be represented by using a single character set.

URL (Uniform Resource Locator):
The 'address' of an Internet-accessible document. Most frequently begins with 'http://…' but also includes 'ftp://…' and 'telnet://…'.

W3C (World Wide Web Consortium):
The organization responsible for managing standards for the WWW.[iii]

XML:
Extensible Markup Language. A non-proprietary text format that is self-describing and flexible, making it attractive as a preservation format. XML is derived from the Standard Generalized Markup Language (SGML).

XHTML:

Extensible Hypertext Markup Language. This information standard essentially expresses HTML code in XML syntax. XHTML 1.0 has been recognized by the Internet-related vendors as the successor to HTML 4.0 and is the equivalent of the most recently adopted HTML 4.1 protocol.

Sources

File Extension Source.
http://www.filext.com/

Society of American Archivists
http://www.archivists.org/glossary/term_details.asp?DefinitionKey=1185.

W3Schools. Refsnes Data.
http://www.w3schools.com/site/site_glossary.asp.

---

[i] Texas Digital Library, *Journal of Digital Information*, Vol. 8, No. 2 (2007), (http://journals.tdl.org/jodi/article/view/229/183).

[ii] W3C, XML Schema 2000

[iii] W3Schools. Refsnes Data. http://www.w3schools.com/site/site_glossary.asp.