COLLABORATIVE ELECTRONIC RECORDS PROJECT:

AN INTRODUCTION AND OVERVIEW

December 2008 Rockefeller Archive Center

TABLE OF CONTENTS

Citation
Adgent, Nancy. (2008). "Collaborative Electronic Records Project: An Introduction and Overview." Sleepy Hollow, NY and Washington DC: The Collaborative
Electronic Records Project.

EXECUTIVE SUMMARY

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 the 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 the SIA and RAC contributed to the CERP's applicability to a wide range of institutions. The SIA is both the institutional archives and the Smithsonian's records manager, serving all of the contributing units. As SIA is a unit within the larger Institution, ownership of accessioned records is not an issue with most accessions. 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 over fourteen years. On the other hand, the 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 One, the focus narrowed to e-mail for several reasons: 1) other projects were addressing website preservation; 2) CERP funding, staffing, and time limitations precluded a comprehensive approach; and 3) e-mail preservation was urgently needed. The enormous quantity of e-mail generated makes better management of digital communication economically advantageous. With the advent of e-mail came a change in organizational roles. In most companies, every employee assigned an e-mail account acts as a file clerk and records manager, and has the ability to create, destroy, mismanage, and improperly use e-mail. Facing regulatory requirements and exposure to lawsuits, companies, particularly non-profits, cannot afford to ignore privacy, security, and rights ownership issues arising from e-mail. Compiling, communicating, and enforcing an organizational e-mail 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 our findings and research, we 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 integ-

rity 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 e-mail 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 e-mail, missing attachments, lack of file order, deteriorating media, obsolete software, and unknown formats. In the course of testing small caches of e-mail, we 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 e-mail to an XML preservation format, and customized an ingest module for depositing e-mail and related metadata into the digital repository, DSpace. Serendipitously, we teamed with the North Carolina State Archives to refine and further test the preservation schema.

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

The CERP Team

Rockefeller Archive Center
Darwin Stapleton, Principal Investigator
Nancy Adgent, Project Archivist

Smithsonian Institution Archives
Riccardo Ferrante, Principal Investigator
Lynda Schmitz Fuhrig, Project Archivist

Consultants Steve Burbeck & Lawry Persaud

[1] Authenticity means a record that is what it purports to be, i.e., includes the e-mail 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.

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 e-mail archiving. Through this summary of our activities, we are sharing our "lessons learned" with a non-technical audience that may include archivists, records donors and depositors, and other interested non-profit institutions. For other CERP publications, updates, and additional information, see http://siarchives.si.edu/cerp. We intend for our products to remain available on the CERP website indefinitely for adoption and modification by any non-profit organization.

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 the RAC did not have information technology staff as did the 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, the 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, the SIA and the 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 e-mail; however, both the 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, the RAC conducted only one visit to each participating depositor. The SIA, due to sharing the same employer as their depositing units, was able to make repeated visits to all contributors.

Who to Interview

Selection of participating staff members at depositing organizations is key. Ideally, the group would include their records manager, information technology manager, operations manager, and at least some of the e-mail creators whose messages are expected to be deposited in the archive. As part of the interview process, an archivist should work with the depositor's management and IT staff to determine which employees' e-mail will be captured, to establish e-mail folder organization structure and naming standards, and to plan a strategy for regularly capturing and transferring selected folders.

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 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
- E-mail not recognized as an official record
- Lack of employee instruction in e-mail 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 e-mail 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 e-mail system or on the e-mail account owner's desktop hard drive and may no longer be accessible
- E-mail retention is dictated by IT storage capacity and backup policy

Results

From the surveys, we 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 our depositor interviews, the RAC and SIA each developed best practices guidance to assist depositors and archivists with e-mail management. Because the RAC does not receive e-mail directly from active e-mail systems in contrast with the SIA which receives e-mail from both obsolete and active e-mail systems, procedures and guidelines were tailored for our different archiving environments. RAC's guidelines address issues and trends that corporate officers and managers of our 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 e-mail 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.

Many small archives and their donors have not trained employees in proper e-mail 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. The SIA's CERP guidelines focus on the unique characteristics of archiving e-mail records since the SIA already had long-established electronic records management policies and procedures in place as well as a records management department.

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: E-mail Records*" and "*E-Mail 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 we acquired enough information to assess the situation we faced, we obtained cooperation from selected depositors in identifying and capturing inactive e-mail for use as testbed material. Although selecting, capturing, and some testing occurred during Phase I, the latter continued into Phase II and our earlier guidelines were revised based on the knowledge gained while transferring, testing, and processing. In accordance with our commitment to the two RAC e-mail testbed depositors, all information that could identify the messages, creators, recipients, or offices of origin would remain confidential. We altered a standard Deed of Gift form into a Testbed Deposit Form, signed by both parties, to reflect our agreements. We 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 will be accessioned by SIA at the conclusion of the project. The remaining material will be destroyed, as the originals remain with the testbeds.

During the transfer process, we 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 we 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 e-mail 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, we 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. We discussed the general contents of various Inbox folders with the two program officers involved, opened a small percentage of messages on their desktops, and determined which folders contained the information we 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 RAC's was viable, thus our appraisal decision to attempt preservation was intuitive.

The three SIA testbeds consisted primarily of Outlook Exchange e-mail accounts and were from administrative, financial, and scientific research units. The first account was transferred as a group of MSG formatted files over SIA's secure server, while the remaining accounts were PST files, transferred directly to a secure server or via ftp server. Accounts were chosen on the basis of imminent departure of two key staff members, the historical research value of another unit's e-mail, and the large number of official records generated by the third unit.

Testing and Processing

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

- Should we 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 do we isolate or remove viruses?
- How will attachments be linked to e-mail messages throughout processing?
- How do we determine if and when native attachments should be migrated to new and/or stable formats
- How do we determine when removable media should be refreshed?

RAC's first processing step was to make two copies of the original, native e-mail being transferred, making three sets – the original, a redundant copy, and a working copy. All

[2] PST refers to Microsoft Outlook's Personal Storage files, a proprietary format that creates one file containing all selected e-mail messages and attachments and stores it outside of the e-mail server.

processing was performed on the working copy. Then we 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. SIA used primarily Outlook and also tried Mozilla Thunderbird. Next we conducted virus scans with the commercial, off-the-shelf anti-virus software used by our respective institutions for non-CERP work. Results of authenticity and integrity verification and virus checks were recorded on the Electronic Records Verification Form at the RAC.

Virus programs differed in their findings and, in one case, the viruses found in e-mails or their attachments could not be quarantined or cleaned when moved from CD to a PC desktop. Luckily, the viruses detected were old (we were testing e-mail created 2 to 6 years prior) and posed no threat to current operating systems. Ideally, e-mail should be cleaned by the depositor before transferring; however, transferred e-mail 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 the RAC's transfers were from external organizations, we 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 e-mail creator, date range of the batch being transferred, type of content (e-mail, spreadsheets, database, etc.), format (Outlook), type media (CD, server, etc.), source (desktop, server, portable device), and the destruction date.[3] The RAC also modified an Accession Form for e-mail. 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 for this documentation.

Another decision facing archives is whether to accept e-mail from depositors who have not deleted personal, sensitive, confidential, and SPAM messages. If unsorted e-mail is accessioned, the archive then has to determine whether to use an archivist's time for this purpose. The 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. Nei-

[3] 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.

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

Attachments pose still another preservation challenge. Because of the variety of attachment file formats, the attachments may obsolesce at a rate different from the e-mail messages. The question facing archives is whether we undertake more time-consuming work to assess their long-term format viability, and potentially extract and migrate them to stable, recommended preservation formats. The SIA decided on that course of action for the testbeds and, after trial and error with various applications, used EZDetach software to remove the attachments, then used JHOVE and DROID along with an SIA-developed tool to aggregate and automate the attachment format analyses. The 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 we expect to accession most of our e-mail more than 3-5 years after creation. Both SIA and RAC kept the original attachments with the source e-mail.

One standard archival decision – whether to impose an order on a collection when no original organization was done – needs to be addressed for e-mail accessions also. This may be the case when the transferred e-mail does not arrive in the context of an e-mail account or the account had only an Inbox folder and no further structure. Depending on the size of the e-mail account being transferred, organizing e-mail 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 we tackled the task of converting e-mail to XML (eXtensible Markup Language), our selected preservation format. Our 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, we investigated software that would convert the original, native e-mail formats into MBOX.[4] Because the RAC's testbed e-mail consisted of a large variety of e-mail 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 e-mail, both the RAC and SIA used Mes-

[4] MBOX is a generic format likely to be viable for decades. For more information, see http://en.wikipedia.org/wiki/Mbox.

sageSave to convert the proprietary PST format into MBOX.[5] Due to the processing time involved and the possibility of mistakenly overlooking recordworthy material, SIA did not sort out messages. The RAC, on the other hand, produced a "researcher use copy" which has had the personal, sensitive, confidential, and junk mail removed.

Both the RAC and the SIA produced finding aids for testbed material, and the SIA created theirs as EAD using NoteTabPro initially, later using oXygen. We adopted a model and workflow process based on the OAIS-Reference Model that starts with a Submission Information Package (SIP) containing the original e-mail transfer and the metadata narrative (i.e., Accession Form). It becomes an Archival Information Package (AIP) with the addition of the finding aid, converted e-mail, parser output, updated metadata, and METS files; then it becomes a Dissemination Information Package (DIP) for retrieval from DSpace. Development of processing workflow and tools continued into Phase III. The workflow is documented on the CERP website.

Results

Throughout the testing phase, we documented changes when particular software was used, what actions were taken, and any anomalies. From this, we 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. We successfully processed e-mail using standard archival concepts of appraisal, accessioning, original order, organization, description, and conservation assessment.

Phase III: Preserving, Storing, and Retrieving

Preserving

Early in the project, we decided we would pursue e-mail 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. XML was chosen as the preservation format because it is human readable, self-describing, its schema supports validation checking, and many open source tools can create, read, and manipulate XML. Dr. Steve Burbeck, our IT consultant, began developing a parser to translate e-mail from the generic format (MBOX) into XML for long-term preservation. Serendipitously, he began talking with David Minor of the North

[5] MessageSave is a software program that converts Outlook e-mail into other formats such as MBOX, MSG, TXT, and EML. For more information, see http://www.techhit.com/messagesave.

Carolina Department of Cultural Resources who had designed an XML schema for use with that state's e-mail account preservation project, EMCAP. Their 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 e-mail preservation.

Collaborating on the schema and testing it on both projects' testbed e-mail 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 e-mail messages, associated metadata, and attachments from MBOX into a single preservation XML file that includes the e-mail 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 e-mail in one XML file, each attachment over 25 KB into a separate XML file, each bad message into a separate file, 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.

Storing

We decided to use DSpace as our testbed digital repository, primarily due to its large user community.[6] Although the preserved e-mail accounts include the Internet Header automatically generated metadata, DSpace requires its own set of metadata for ingest into the digital repository. Our second IT consultant developed an ingest module that uses Metadata Encoding Transmission Standard (METS). We selected a set of ten key metadata elements to describe the AIP (including Record Group, Account Holder, Date Range) required for deposit in DSpace. The AIP Metadata for E-Mail Form designed correlates each element with DSpace terminology and the Dublin Core tags required by the ingest utility. The METS file is completed manually.

The AIP stored in DSpace consists of:

- source format e-mail account (.pst, .msg, etc.)
- MBOX format e-mail account (preliminary preservation transformation)

[6] 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/.

- 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

The 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, our recommendation is to store another, redundant copy offsite in a proper physical environment. SIA retains their original and preserved accounts online with redundant copies on an external drive and tape.

Retrieving

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 e-mail?
- 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 e-mail 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 Library of Congress or other standards?

Results

We achieved a 99+% 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. 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 e-mail messages themselves.[7] We deposited testbed e-mail accounts into DSpace using our METS ingest module and were 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. We would like to see our 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 use 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
8. Retention/Destruction: Automate a calendar-based notification system that would alert an archivist when particular files need to be migrated, refreshed, or destroyed.

Sharing Our Experiences

Presentations

Throughout the project, team members shared findings with other archivists and groups interested in archiving e-mail. Questions asked at these events helped shape our guidance documents, outreach efforts, and ultimate effectiveness. Also informal conversations while attending seminars and conferences kept our peers informed about our progress

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

and gave us additional ideas for content of project deliverables. Both the RAC and SIA regularly updated our staffs and addressed small groups of employees and others upon request.

During the first phase (September 2005-December 2006), we accepted invitations to give formal presentations at a broad range of venues: the American Institute for Conservation of Historic and Artistic Works annual meeting; the Westchester County Historical Society and the Westchester County Archives conference; and the Archivists' Roundtable of Metropolitan New York. We also made presentations to the RAC's new Board and to a visiting committee assessing CERP's value to the archival community. The following year, at the request of an SAA Advanced Electronic Records workshop leader, we conducted a two-hour segment summarizing our work to date, distributed guidance documents and documentation forms, and answered questions. SAA accepted our proposal for the poster session at the August 2007 conference where we attracted considerable interest, discussed particular issues with many attendees, and distributed brochures.

In the final year, conference program committees expressed more interest in CERP and our presentation proposals were accepted by the Midwest Archives Conference, Society of North Carolina Archivists, Southern Archives Conference, Australian Society of Archivists, and Society of American Archivists conferences and the SAA Research Forum. We conducted a day-long workshop at the Association of Canadian Archivists conference that included hands-on exercises and demonstrations using various software to process, convert, and validate e-mail. SIA CERP members used their proximity to other large institutions such as the Library of Congress, the National Archives and Records Administration, and Congressional offices to further disseminate our findings and products. Throughout the project, RAC visitors received a CERP overview and handouts as part of their orientation and tour. Near the end of the project, we held a final symposium at which future development possibilities were discussed.

Publications

In addition to compiling project brochures and guidance documents, we wrote several articles that were published in RAC newsletters, in SAA Section newsletters (College and University, Electronic Records, Government Records, Manuscript Repositories, and Preservation), and on the Museum Archives Section blog. We initiated a "*Friends of CERP*" electronic newsletter in December 2006. At seminars and conferences we distributed CERP brochures and newsletter articles.

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 us to locate IT consultants who could accomplish the tasks of developing a parser and customizing DSpace ingest. We were 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 standards such as EAD, DACS, XML, 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 e-mail, 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 serendipitously simply by following up on conversational twists and by seeing the Inbox organization and the e-mail 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 e-mail 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 e-mail from particular depositors.
- Progress is hampered if using outdated computers and ones with inadequate RAM.
- Testing as wide a range of e-mail 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 e-mail 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 e-mail expected – unless automated tools are developed.

Preserving, Storing, and Retrieving

- No parser will work on 100% of the e-mail ever created.
- 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 our Experiences

- Our experience does not provide answers to every problem.
- Archivists, donors, and publishers of archival literature are thirsting for information about e-mail 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 e-mail 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 your 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 e-mail 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 e-mail 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 e-mail 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/.

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.[1]

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.

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.

Electronic Communications Privacy Act (ECPA):
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.

Enterprise Content Management (ECM):
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-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.

Open Archives Initiative (OAI):
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.

Portable Document Format (PDF):
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.

Personal Storage File (PST):
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 Application (RMA) or Records Management System (RMS):
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.

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

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.[2]

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.

Submission Information Package (SIP):
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.[3]

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.

---

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

[2] W3C, XML Schema 2000

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