

Digital Dilemmas: Preserving E-Mail

Smithsonian Institution Archives
and Rockefeller Archive Center
tackle the challenge of archiving
e-mail and attachments.



Smithsonian Institution Archives

THE ROCKEFELLER ARCHIVE CENTER

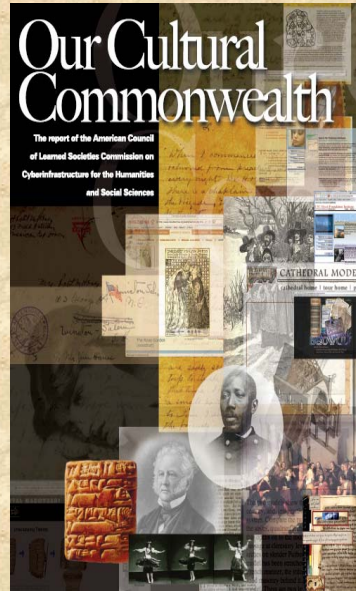
September 13, 2007, Smithsonian Institution

Collaborative Electronic Records Project (CERP) presented by Lynda Schmitz
Fuhrig, Smithsonian Institution Archives project archivist.

Our digital world

“In 2006 most expressions of human creativity in the United States – writing, imaging, music – will be ‘born digital.’ ”

“In addition to digitizing materials, projects to collect and preserve born-digital content are critically important.”



Email and other electronic files present an abundance of digital dilemmas to today's records managers and archivists.

“Our Cultural Commonwealth,” a 2006 report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences, had this to say about born-digital material.

See <http://www.acls.org/cyberinfrastructure/OurCulturalCommonwealth.pdf>

Collaborative Electronic Records Project

- Testbed surveys
- Email guidance
- Transfer guidelines
- Long-term preservation of email, attachments, & other digital files
- Digital repository



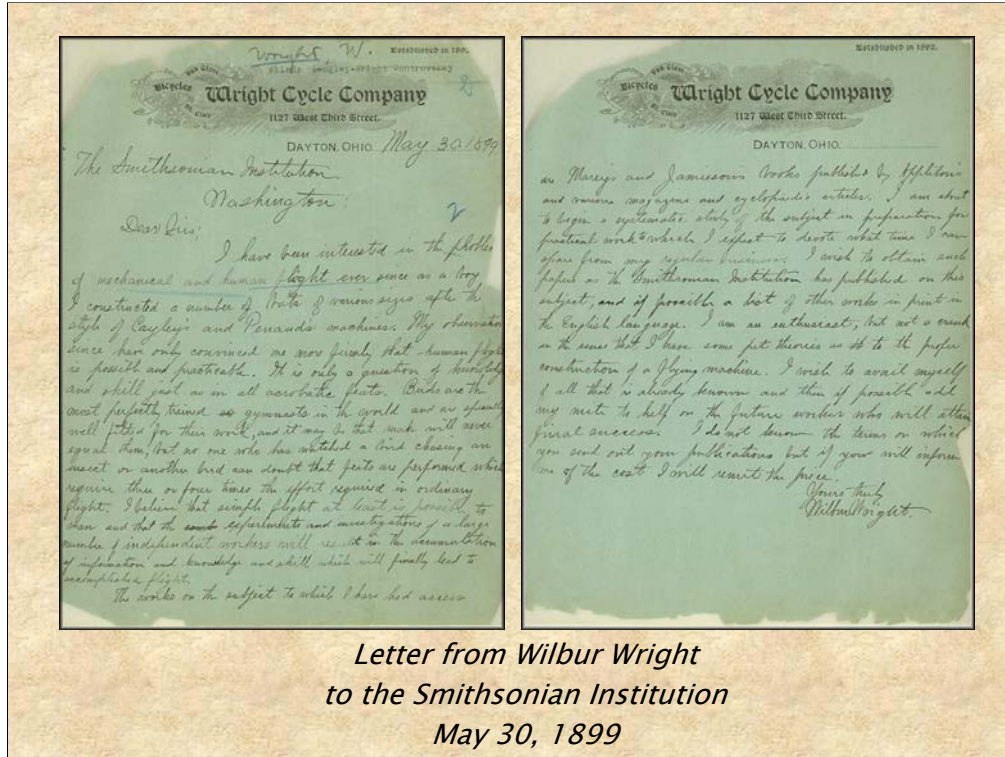
The Rockefeller Archive Center and the Smithsonian Institution Archives have depositors that rely on email, databases, spreadsheets, word processing, design programs, and other software packages to carry out the functions of their organizations. SIA and The Rockefeller Archive Center decided to explore long-term preservation of born-digital materials, primarily email messages and attachments because of the prominent role email plays in today's organizations. We are now in the final year of the project.

Pictured is Rockefeller Archive Center project archivist Nancy Adgent speaking to an attendee at the Society of American Archivists 2007 conference in Chicago at our project poster.

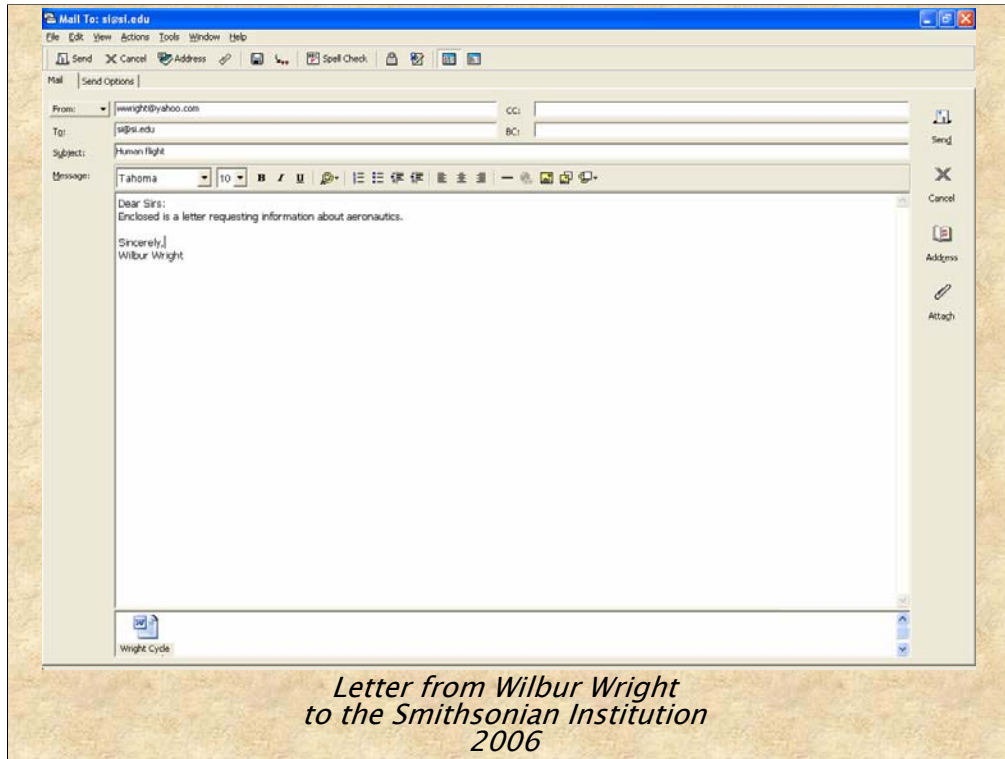
We talked with some of our colleagues at other institutions during SAA about email. Some comments we heard were that they are just starting to tackle email:

- A professor is retiring and planning to transfer his email records and they are not sure how to do this.
- Others have been printing out email messages and filing into folders.
- Or they having been saving the "important" emails as PDF files.

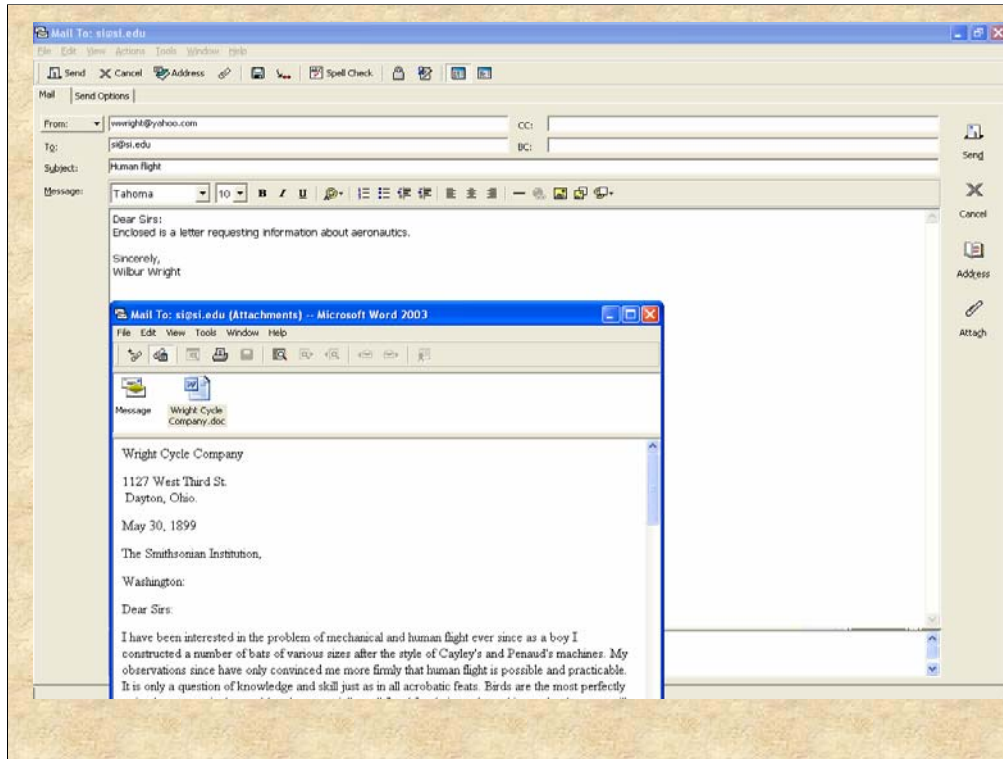
Are these the best solutions?



Handwritten letters have provided scholars with tremendous insight into the everyday lives of notable and not-so notable figures; today's email messages have taken on that role.



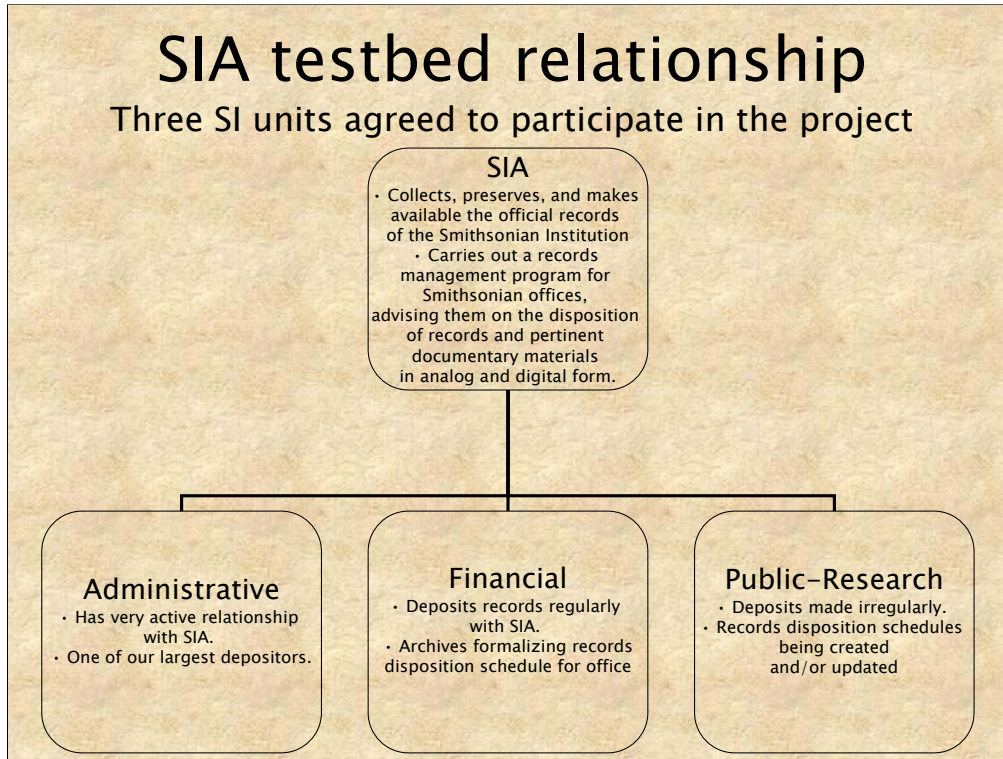
Wilbur Wright's letter to the Smithsonian might be written in Word and sent as an email attachment today. Obviously, it's not as visually interesting as the handwritten document.



Email messages can contain vital information that makes them recordworthy. Email also can provide a snapshot of the inner-workings of an organization – a chance to view relationships, hierarchies, and social behavior.

Of course email has its challenges. One SI employee made the observation that email messaging is a platform that contains important business decisions and right next to it is a message about Viagra.

This summer a report was issued by an independent review committee about SI that contained internal email messages. Needless to say, SI email records from early in this century will provide some fascinating materials for researchers decades from now.



Three units at SI agreed to participate as testbeds in the first phase of the project. I interviewed nearly 40 staff members, compiled information and selected employees who appeared to be good candidates for providing copies of email, attachments, and other digital documents. I learned about the missions and histories of the offices, recordkeeping practices in digital and analog form, and general operations.

With consultation with our Records Management (RM) Team, we recommended potential records for capture. More follow-up interviews were conducted with selected staff about their email practices, i.e., arrangement of email within the application, how long messages are retained, and attachment practices. At the end of the project the Records Management Team will decide which records it will accession into the archives for permanent retention.

One unit dropped out about halfway into the project due to time issues. Nevertheless, we still have good material from this office to use for CERP and have continued working with them in other ways.

What has been taken in for the project

SIA

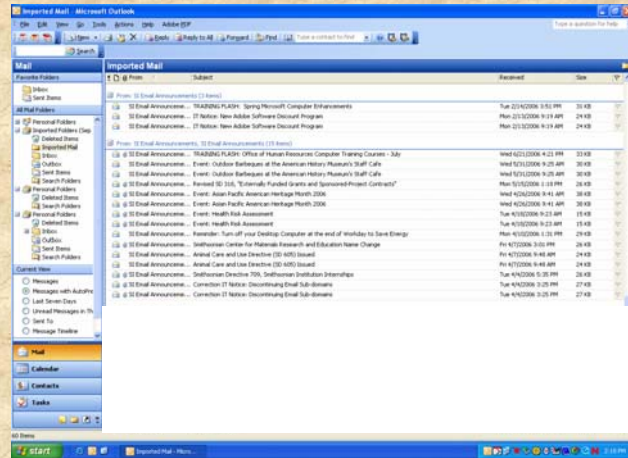
- 2.7 GB of email from three units or 36,026 messages
- Other digital material includes program documents such as images, video, and manuals. Formats include MS Office Suite, WordPerfect, PDF, JPEG, GIF, and TIF

RAC

- 817 MB of email from two units
- Other material from scanning project with now obsolete software

Challenges

- Proprietary formats
- Depositor practices vary (Subfolders, Inbox only, etc)
- Renderability problems
- Personal with the business records
- Manual and time-consuming processing
- Confidentiality issues



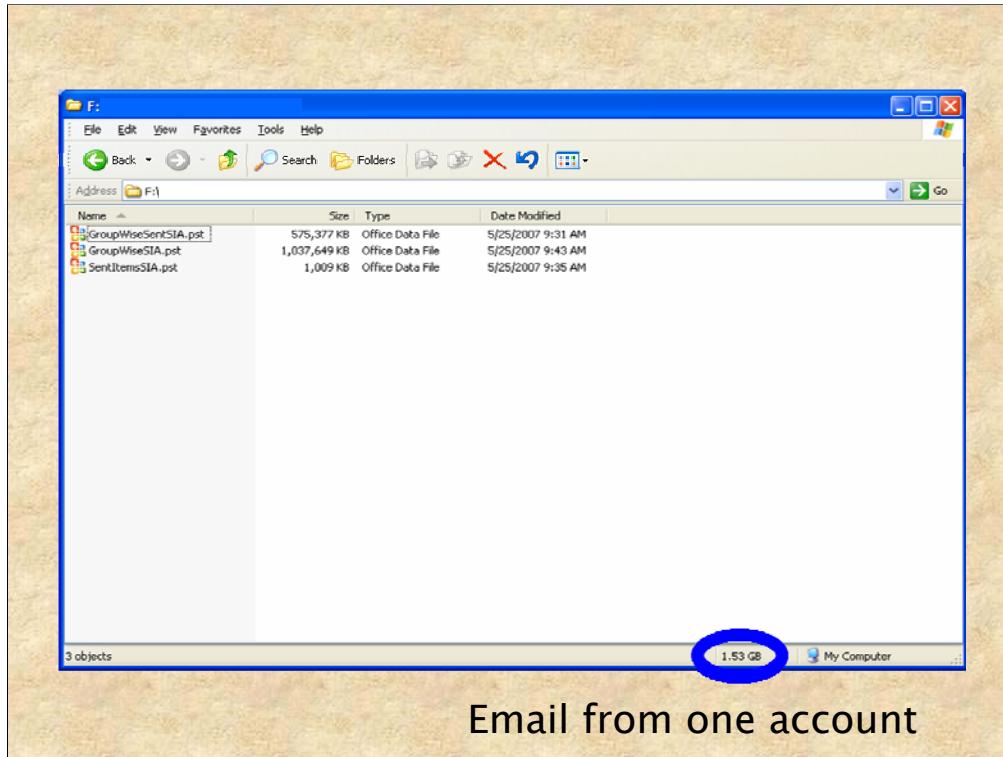
The transfer process involved capturing accounts within a certain date range and certain folders (when used) within MS Outlook/Exchange. These messages were copied into a .pst file and retrieved. (.Pst stands for Personal Storage or Personal Stores, which stores email and attachments outside of the email server. It essentially creates one file of all the email messages and attachments saved to it. This file can only be opened in Outlook but the messages are viewable as separate emails). SI is migrating all units to Outlook for email.

The first group of email was a server transfer done by the recipient. She was asked to sort out specific email with keywords agreed upon with our RM Team. We later decided the keyword method was not feasible because too much recordworthy material could be missed. She was not able to create the .pst file and sent separate msg files, which were converted to the .pst format using a tool called Aid4Mail.

In later captures of email, we asked recipients to weed their own accounts. This gave them the opportunity to clean out non-business or outdated material. This method is not perfect either because non-business or non-essential emails remained in some accounts, such as news alerts from CNN and restaurant reservations.

We went on site to retrieve these .pst files for transfer to our server with these remaining accounts. The manual process of transferring can take anywhere from 30 minutes to 90 minutes. Another retrieval method involved a tool from Microsoft called ExMerge but it pulled the wrong dates.

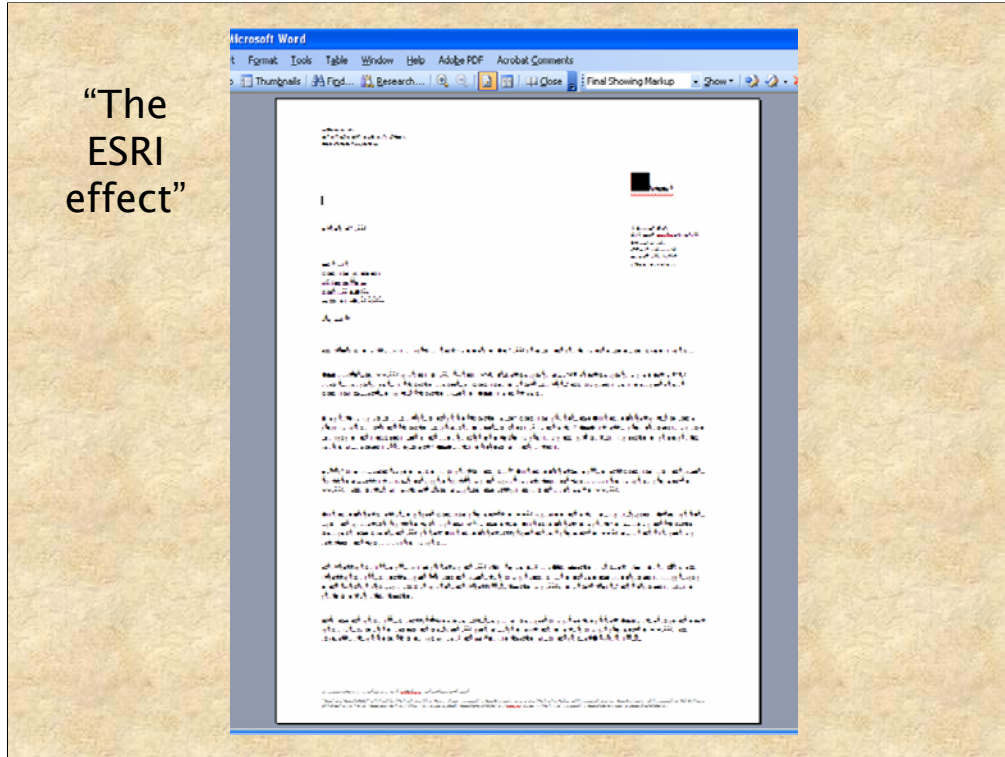
Speaking of .pst files: They are proprietary and they can become corrupt when they get large. There is a 2GB limit.



Practices vary among email users. Some just have an Inbox with more than 10,000 email messages. This account is 1.5 GB with more than 20,000 items. Another account has 100 subfolders and folders within those folders.

“Processing” is time consuming and manual. How do you deal with 14,000 items in a Sent Items folder? Think of someone who has folders/boxes filled with papers with no organization piled to the ceiling in their office. While item-level processing is unrealistic with large accounts, it has been useful with smaller accounts to get an idea of subject materials, senders, and relationships. This also demonstrates patterns or items to be aware of in future email accessions.

“The ESRI effect”



Renderability – The ESRI effect. This is from an email attachment. When opening the Microsoft Word document on my PC, this is how it appeared. I highlighted the text and selected another font but it clearly was not the correct display. I also opened it in OpenOffice, which is open-source office suite software. The document displayed in readable text. Finally, we decided to open the document in NotePad to find out about the Word fonts. We discovered the format conversion within MS Word was set on my PC to ESRI fonts, which are from the GIS software. I had ArcGIS installed on my PC in 2006. There is an attachment also within this attachment, which wasn't apparent in this view.

Amazing animal tricks and other issues



Tyson the skateboarding bulldog

This is a screenshot from a video of a skateboarding dog named Tyson. The video was another email attachment that came from a colleague at another organization in 2004, before the iPhone commercial. The recipient was blind carbon copied on the message. A few months later the recipient replied to that same email with a professional inquiry. She retained the original subject line, which had nothing to do with the business-related question. The respondent also kept the same subject line. If someone is looking for the business-related email message and only browsing/searching subject lines, it could be missed because it is labeled “skateboarding dog.”

Confidential information is also a problem in email correspondence. It can be easy to forget information such as Social Security numbers is in an attachment, especially if it is embedded in the message of a message.

See <http://www.youtube.com/watch?v=QaQw9V4Upj4> for the video
Tyson information at <http://www.skateboardingbulldog.com/>

Outputs

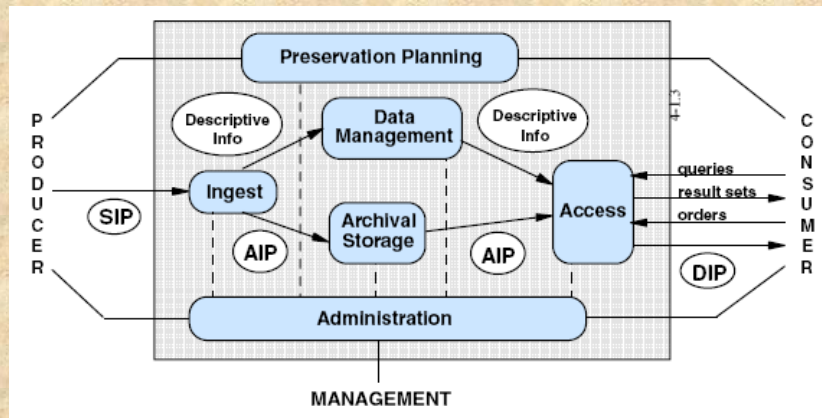
- Recordkeeping and email guidance documents
- Transfer guidelines
- Website and quarterly bulletin
- Completing and/or revising records disposition schedules for two offices



From our work so far we have issued transfer documentation for the digital files, recordkeeping and email guidance, including information on weeding, what makes an email message a record, and examples of poor email management consequences, as well as offering assistance to email users on search capabilities and .pst creation.

SI Archives also is completing and/or revising records disposition schedules for two of the testbeds.

OAIS



Consultative Committee for Space Data Systems

Open Archival Information System functional model

We are using the OAIS (Open Archival Information System) functional model from the space data community as our framework for a digital repository system. It 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 this framework for their digital preservation efforts because of its flexibility and acceptance. Following this model helps ensure the integrity and authenticity of the contents and system throughout the lifecycle.

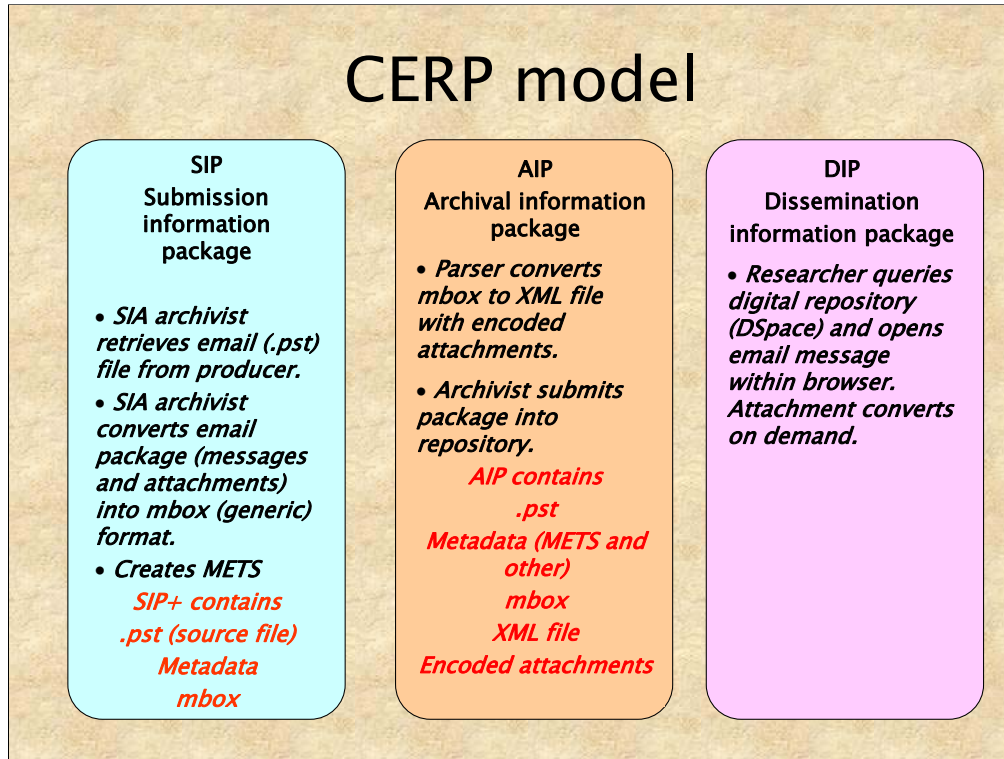
SIP – Submission Information Package

AIP – Archival Information Package

DIP – Dissemination Information Package

See <http://public.ccsds.org/publications/archive/650x0b1.pdf>

CERP model



This is the CERP model draft. The email package with metadata is received and processed. Checksums are to be applied upon ingest. Processing can include an overview of the messages, virus check, and subject line and attachment extraction. The .pst is converted into a generic mailbox format with a tool called MessageSave and then is transformed into XML with the encoded attachments. These files (.pst, generic file included) are placed into a digital repository, and retrieved when requested. Attachments are converted on demand.

We are continuing our testing with the XML conversion using a schema developed in collaboration between the CERP technical consultant and the Electronic Mail Preservation Collaboration Initiative (EMCAP), a joint project with North Carolina, Pennsylvania, and Kentucky. See <http://www.ah.dcr.state.nc.us/records/EmailPreservation/default.htm>

XML is not proprietary. It can create HTML, PDF, and word processing documents all from the same XML.

The formats for the project have been mostly straightforward but our Institution will continue to deal with a growing body of new and complicated software applications. Most archives, SIA included, have received digital files that are unreadable due to the media or software used to create it. Needless to say metadata is crucial for future access. We are exploring METS, Dublin Core, and PREMIS.

See <http://www.loc.gov/standards/mets/>

<http://dublincore.org/>

<http://www.loc.gov/standards/premis/>

What's next



DSpace testing

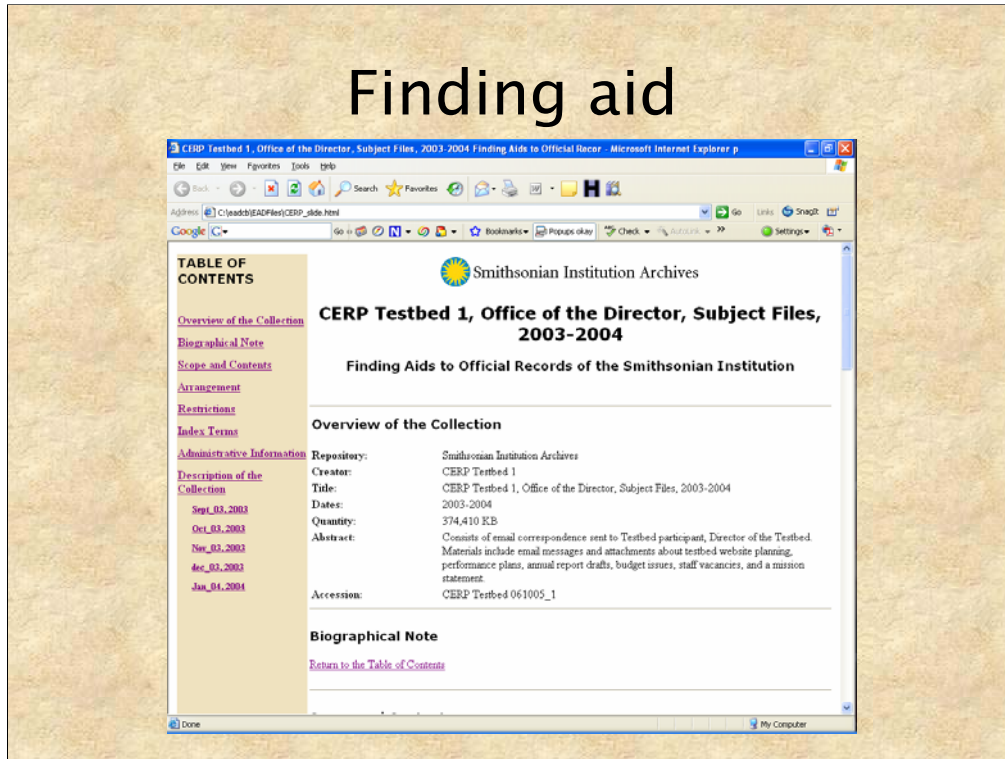
One of CERP's goals is to have digital repository model based on affordable, open-source tools so small to mid-sized institutions can adopt/adapt it for their needs. COTS or commercial-off-the-shelf systems are pricey, proprietary, and do not focus on long-term preservation.

We plan on using DSpace for the digital repository testing for this project. A number of academic institutions and some government bodies, including the state of Kansas, are using this repository software for digital material. This open-source application developed by MIT and Hewlett-Packard has a strong community and recently started the DSpace Foundation, ensuring a commitment to longevity. Smithsonian Institution Libraries has been using DSpace since late 2005. The Smithsonian Digital Document Repository collects digital SI scientific publications.

See <http://www.dspace.org/>

<http://si-pddr.si.edu/dspace/>

Finding aid



We also want to create finding aids for this project. I have started exploring EAD (Encoded Archival Description). It is a bit challenging at this stage with email accounts. Archives of American Art has generously shared information with us about the tools it uses.

We also plan to hold symposia in 2008, here in Washington, D.C., and at the Rockefeller Archive Center in Sleepy Hollow, N.Y.

See <http://www.loc.gov/ead/>

<http://www.aaa.si.edu/>

For More Information

CERP website

<http://siarchives.si.edu/cerp/cerpindex.htm>

Smithsonian Institution Archives website

<http://siarchives.si.edu/>

Rockefeller Archive Center website

<http://archive.rockefeller.edu/>

Lynda Schmitz Fuhrig

schmitzfuhrigl@si.edu

Digital documents can be fragile just like paper but in a different way. The sooner we can accession them the better. Good records management for depositors is an ongoing issue of education. We need to accept that paper is not going to go away, but there should not be a need to print out every email for posterity, nor is it feasible. As we move into our last phase, we hope to provide some answers to this digital dilemma.