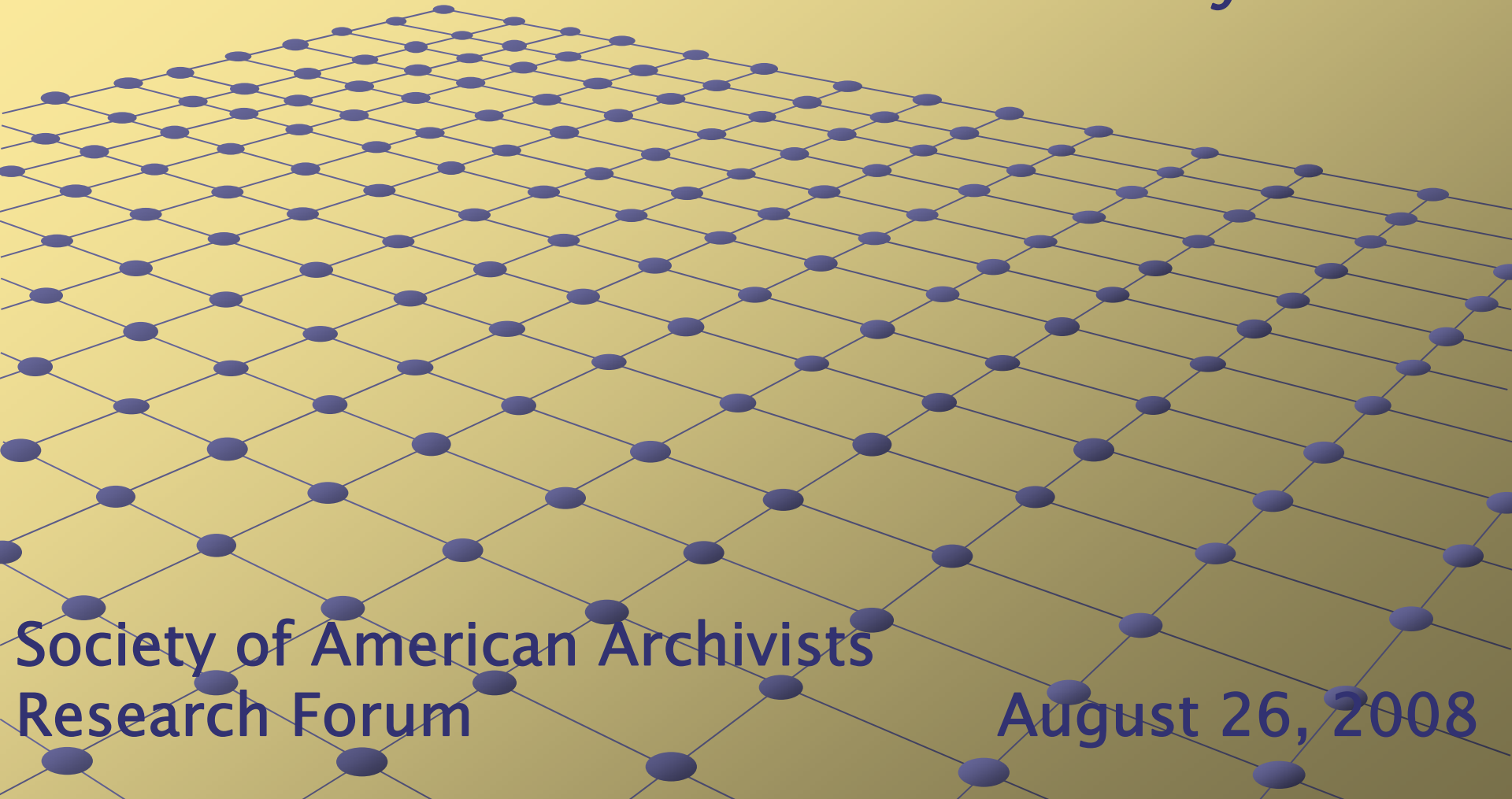


Digital Dilemmas: Archiving E-Mail

Collaborative Electronic Records Project



Society of American Archivists
Research Forum

August 26, 2008





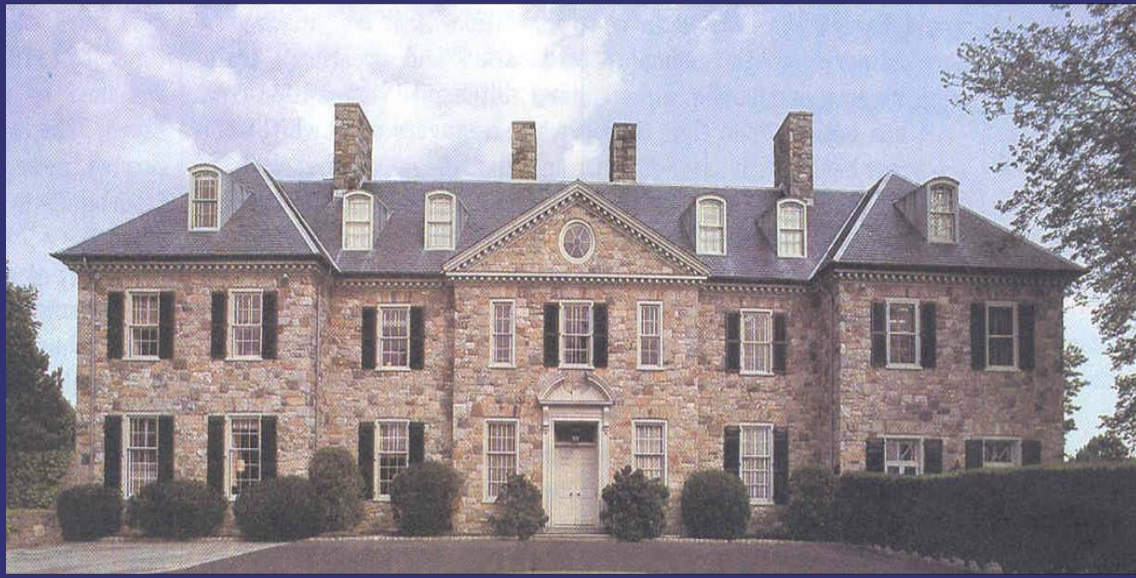
Nancy Adgent

Project Archivist
Rockefeller Archive Center

15 Dayton Avenue
Sleepy Hollow, NY 10591

914-366-6355

nadgent@rockarch.org



Rockefeller
Archive
Center
(RAC)

Smithsonian
Institution Archives
(SIA)





Key Survey Findings

- No records management policy
- No naming standards
- No procedures for organizing or saving
- Some have no on-site IT staff



ISSUES

- Unknown formats
- Deteriorating media
- Data on portable devices
- Native format vs. converting
- Upgraded hardware/old media
- Obsolete or unsupported software
- Duplicates, personal, junk mingled
- Information quantity & rate increase
- Traditional archival concepts/new era

Best Practices Guidance



E-MAIL GUIDELINES

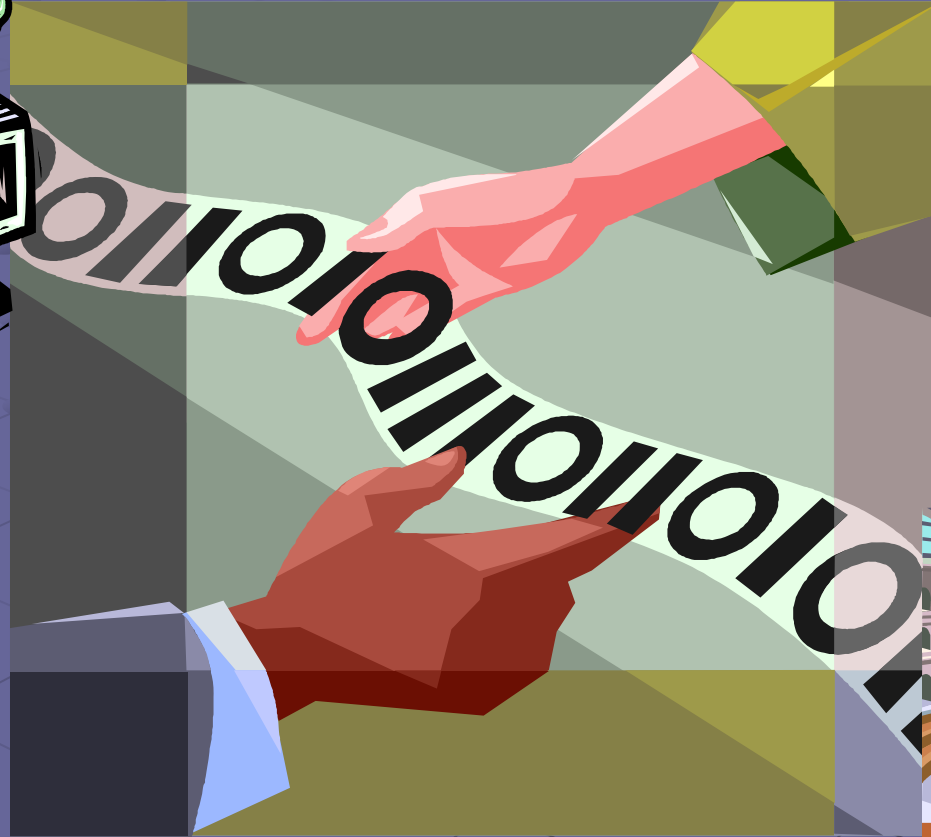
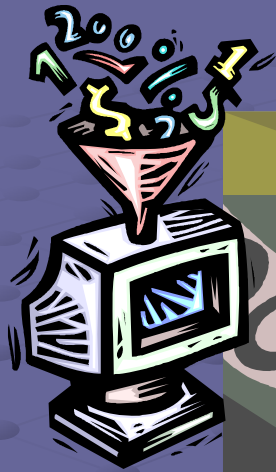
RECORDS RETENTION AND DISPOSITION GUIDELINES

Prepared by the Collaborative Electronic Records Project
Rockefeller Archive Center
October 2007

This document may be freely used and modified by any non-profit organization.



TRANSFER GUIDELINES



Prepared by the Collaborative Electronic Records Project
Rockefeller Archive Center
January 2007

This document may be freely used and modified by any non-profit organization.



Forms

- Accession Administrative & Descriptive Metadata
- Transfer
- Verification
- Migration/Refresh
- METS AIP Metadata

Accession Administrative Metadata

ACCESSION ADMINISTRATIVE METADATA for E-MAIL <METSamd>

Accessioning Archivist: Nancy Adgent
<dc:contributor>

Today's Date: 2008-02-19
<dc:date.created>

Depositing organization <dc:rightsholder>: Rockefeller University
Contact name: Jane Grant Administrator Doe
Address: 1230 York Avenue, New York, NY 10065
Telephone: 212-327-7900
E-mail:

Type of Accession <dc:accrualMethod>: Deposit, Testbed, temporary

Type of Material <dc:type>: Email/mbox with attachments

Size <dc:format.extent>: 551.4 MB

Terms/Restrictions <dc:rights>: Closed to researchers

Access <dc:accessRights>: CERP team only

Retain until <dc:accrualPolicy>: 2008-09-01

Copyright: No (If yes, describe)



Collection: Rockefeller Archive Center

Electronic Records Verification Form

Collection: Rockefeller Archive Center

Electronic Records Media Refresh/Migrate/Destroy Schedule

[illegible]



TESTBEDS

● RAC

- 1) Outlook .pst files from server
- 2) Variety of native e-mail clients from desktop

● SIA

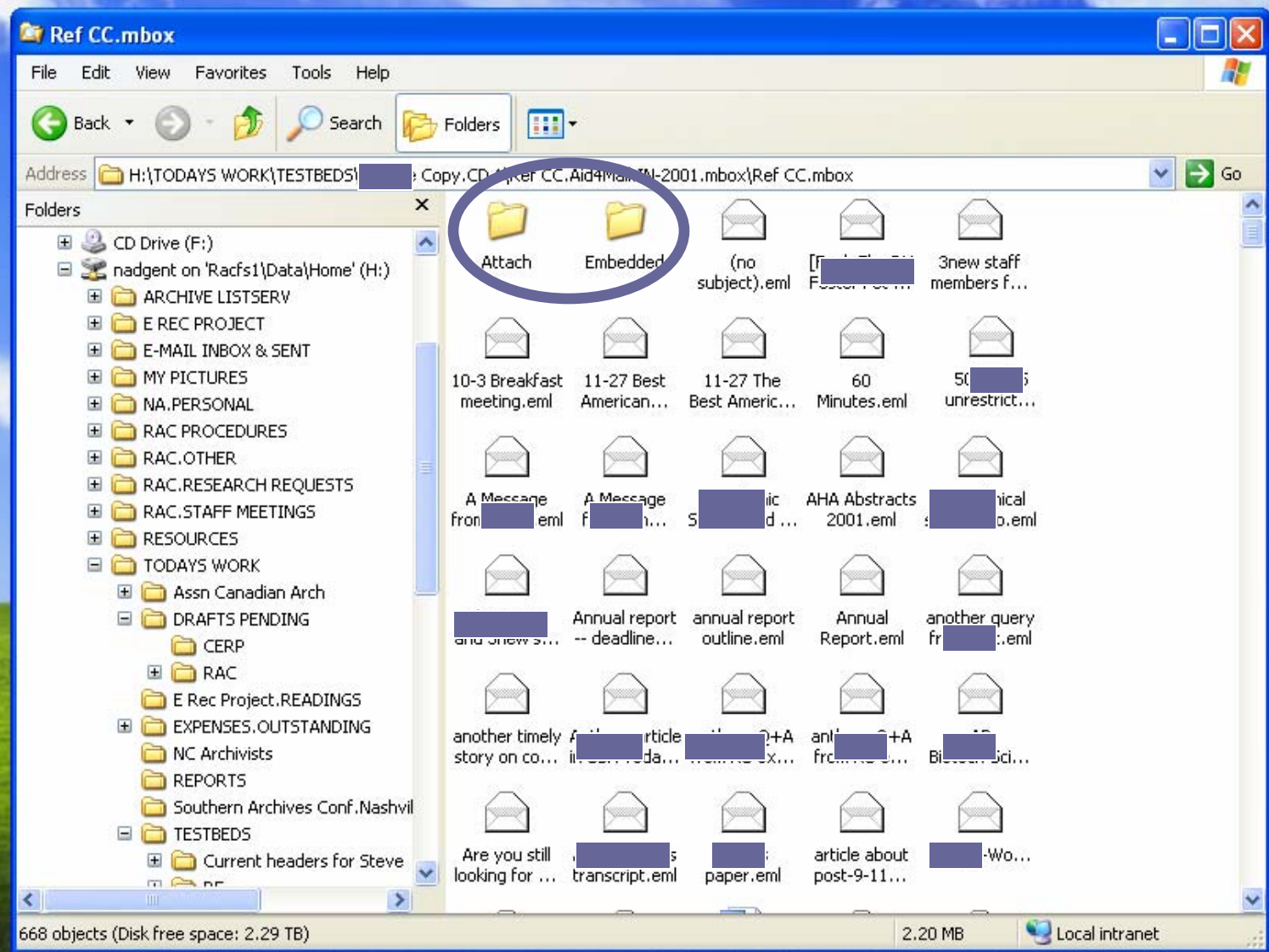
- Outlook .pst files from active system



Web Browser Display

```
-----1211362437=====E2mXatt
Content-Disposition attachment: filename="XXXXX.doc"
Content-Type: application/octet-stream; name="XXXXX.doc"Content-Transfer-
Encoding: base64
OM8R4KGxGuEAAAAAAAAAAAAAAAAAAAAAAAAAAPgADAP7/CQAGAAAAAAAAAAAAAAAAABAAAAJQ
AAAAAEAAAJwAAAAEAAAD+////AAAAACQAAAD////////////////////////////////////
14 pages of character strings were in this space.
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA--
-----1211362437=====
From ???@??? Mon Sep 17 16:44:44 2001
Return-Path: <XXXXXX@mail.rockefeller.edu>
Received: from [123.45.67.890] (XXXXXX.rockefeller.edu [123.45.67.890]) by
mail.rockefeller.edu (6.23.5/7.89.0) with ESMTP id f8HK8Js01021 for <XXXXXX>;
Mon, 17 Sep 2001 16:08:19 -0400 (EDT)Message-Id:
<a0501040bb7cc1683f959@[123.45.67.890]>Date: Mon, 17 Sep 2001 16:07:46 -0500To:
XXXXXXXXXXFrom: Jane Doe <jdoe@mail.rockefeller.edu>Subject: Edited version of letterX-
UIDL: ?e%!!Z"H!!V=^!!+[~!!Mime-Version: 1.0Content-Type: multipart/mixed;
boundary="-----1211361629=====This is a multi-part
message in MIME format.-----
1211361629=====Content-Type: text/plain; charset="iso-8859-1"--
-----1211361629=====Content-Type:
text/plain;"XXXXXXXXX.doc 1 (missing attachment)-----
1211361629=====
```

Aid4Mail Conversion





TOOLS

● MessageSave

● EZDetach

● Aid4Mail

● Fentun

● JHOVE

● CERP Parser

● DROID

● DSpace



CERP Model

SIP *

SIP to AIP

- Archivist converts the collection to the .mbox (generic email format), if not already in this format.
- Archivist runs the parser to convert the .mbox file/s to an XML preservation file with encoded attachments.
- Archivist creates a package of all components (metadata, source, outputs, finding aids) in the zip format and submits to a digital repository.

** 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.*

AIP *

** 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).*

AIP to DIP

The researcher queries the digital repository (DSpace) to find and retrieve the email collection results.

DIP *



XML Preservation Format

- Good prospects for format longevity -- Base is ASCII
- Human readable and “self describing”
- Good descriptive schema supports validity checking
- Many open source tools to create, manipulate, and read XML



Importance of a Common Schema

- Defines how XML tags relate to each other
<Account>, <Folder>, <Message>, <Header>, <Body>, <Attachment>
- Rosetta stone that guides how raw email is converted to XML
- Defines the structure for search, display, provenance, preservation, etc.



CERP's Email Account Schema

- Serves CERP and EMCAP purposes
- Supports email accounts from different systems
 - CERP Parser – multiple formats, no original systems
 - EMCAP parser – single format, active original systems
- Final schema fully addresses a complete email account at all levels
- Enables validation
- Will be made public



Email Conversion Results

- Converted and validated 70,000 messages to the XML Mail–Account schema
 - Smithsonian – 5,537 messages in 232 Mb of recent Outlook mail
 - 99.97% successfully parsed (4 could not be parsed)
 - Smithsonian – 28,000 messages in a 1.5 Gb Outlook account
 - 99.975% successfully parsed (5 could not be parsed)
 - Rockefeller Archives – 43,778 messages in 378 Mb of older eclectic mail
 - 99.85% successfully parsed (74 unparsed, but improvement is clearly possible)



Lynda Schmitz Fuhrig

Project Archivist
Smithsonian Institution Archives

Capital Gallery Building
600 Maryland Avenue, SW, Suite 3000
Washington, D.C. 20024-2520
202-633-5917

SchmitzfuhrigL@si.edu

The Parser Web Interface

Seaside - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Media

Address http://localhost:9092/seaside/EmailParsing?_s=SMetNL5sUbuZXuaw&_k=etELxFxJ Go Links »

New Session Configure Toggle Halos Profiler Memory Terminate XHTML 4/17 ms

CERP Email Parsing

The root directory should contain all Account directories to be parsed. Ensure that the root directory full path is correct. If not, it will default to C:\

Then choose the account directory you wish to parse from the drop-down list of candidates available in the root directory. Within an account directory, email must be contained in folders (subdirectories). The email must be, in "mbox" format in files named "messages.mbox", one such file per folder. The account directory must contain all folder subdirectories that you wish to parse. Examples might include Inbox and/or Sent folders. Any folder may itself contain subdirectories representing sub-folders.

Once you have chosen the desired target account, press the "Proceed with parsing" button. If that account has already been parsed, you will be asked whether or not you wish to reparse it.

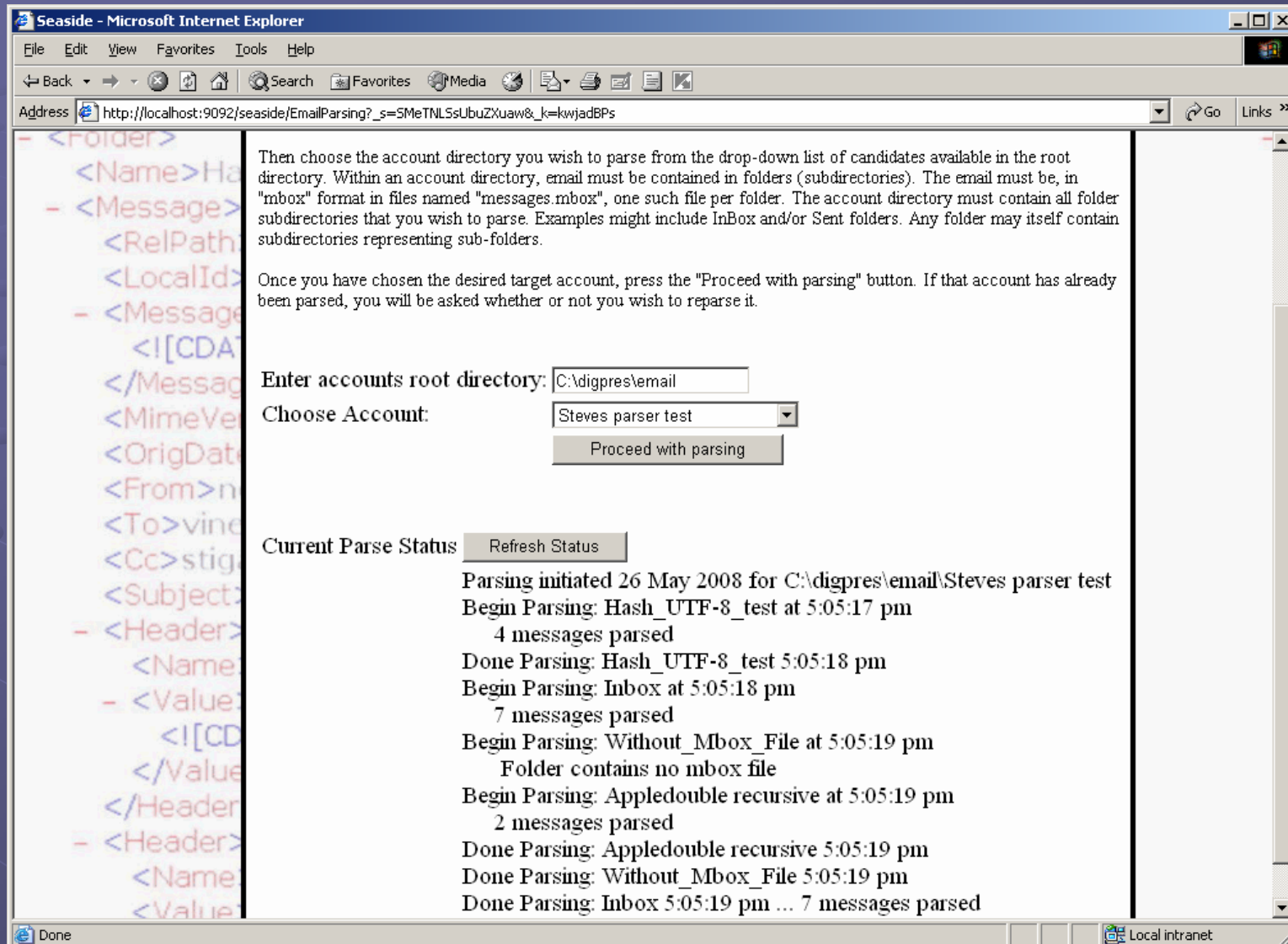
Enter accounts root directory:

Choose Account:

Current Parse Status

No parse status available

Parsing Results Status



Seaside - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Reload Home Search Favorites Media Print Copy Paste

Address http://localhost:9092/seaside/EmailParsing?_s=SMetNL5sUbuZXuaw&_k=kwjadBPs Go Links >>

- <Folder>
 <Name>Ha
- <Message>
 <RelPath
 <LocalId
- <Message
 <I[CDAT
 </Message
 <MimeVer
 <OrigDat
 <From>n
 <To>vine
 <Cc>stigi
 <Subject>
- <Header>
 <Name
- <Value
 <I[CD
 </Value
 </Header
- <Header>
 <Name
 <Value

Then choose the account directory you wish to parse from the drop-down list of candidates available in the root directory. Within an account directory, email must be contained in folders (subdirectories). The email must be, in "mbox" format in files named "messages.mbox", one such file per folder. The account directory must contain all folder subdirectories that you wish to parse. Examples might include Inbox and/or Sent folders. Any folder may itself contain subdirectories representing sub-folders.

Once you have chosen the desired target account, press the "Proceed with parsing" button. If that account has already been parsed, you will be asked whether or not you wish to reparse it.

Enter accounts root directory:

Choose Account:

Current Parse Status

Parsing initiated 26 May 2008 for C:\digpres\email\Steves parser test
Begin Parsing: Hash_UTF-8_test at 5:05:17 pm
4 messages parsed
Done Parsing: Hash_UTF-8_test 5:05:18 pm
Begin Parsing: Inbox at 5:05:18 pm
7 messages parsed
Begin Parsing: Without_Mbox_File at 5:05:19 pm
Folder contains no mbox file
Begin Parsing: Appledouble recursive at 5:05:19 pm
2 messages parsed
Done Parsing: Appledouble recursive 5:05:19 pm
Done Parsing: Without_Mbox_File 5:05:19 pm
Done Parsing: Inbox 5:05:19 pm ... 7 messages parsed

Done Local intranet

Parsed Email Body Excerpt

```
<Value>RO</Value>
</Header>
- <MultiBody>
  <ContentType>multipart/mixed</ContentType>
  <BoundaryString>-----_NextPart_000_0013_01C65275.ED5E9D90</BoundaryString>
  <Preamble>This is a multi-part message in MIME format.</Preamble>
  - <MultiBody>
    <ContentType>multipart/alternative</ContentType>
    <BoundaryString>-----_NextPart_000_0013_01C65275.ED5E9D90_A</BoundaryString>
    - <SingleBody>
      <ContentType>text/plain</ContentType>
      <Charset>us-ascii</Charset>
      <TransferEncoding>7bit</TransferEncoding>
      - <BodyContent>
        <Content>Nancy - Dr. Stapleton asked me to make a few small changes to the draft of the Testbed Agreement
        revised draft (WordPerfect) for your review. I am giving him several copies to take with him for the team meeting
        on Thursday, since he said it would be good to have it for both meetings. Have a good trip. Ken</Content>
      </BodyContent>
    </SingleBody>
    - <SingleBody>
      <ContentType>text/html</ContentType>
      <Charset>us-ascii</Charset>
      <TransferEncoding>quoted-printable</TransferEncoding>
      - <BodyContent>
      - <Content>
        <![CDATA[ <html xmlns:o=3D"urn:schemas-microsoft-com:office:office" =
        xmlns:w=3D"urn:schemas-microsoft-com:office:word" xmlns:st1=3D"urn:schemas-microsoft-com:office:smarttags" xmlns=3D"http://www.w3.org/TR/REC-html40">
```




Parsed E-Mail Attachment Reference

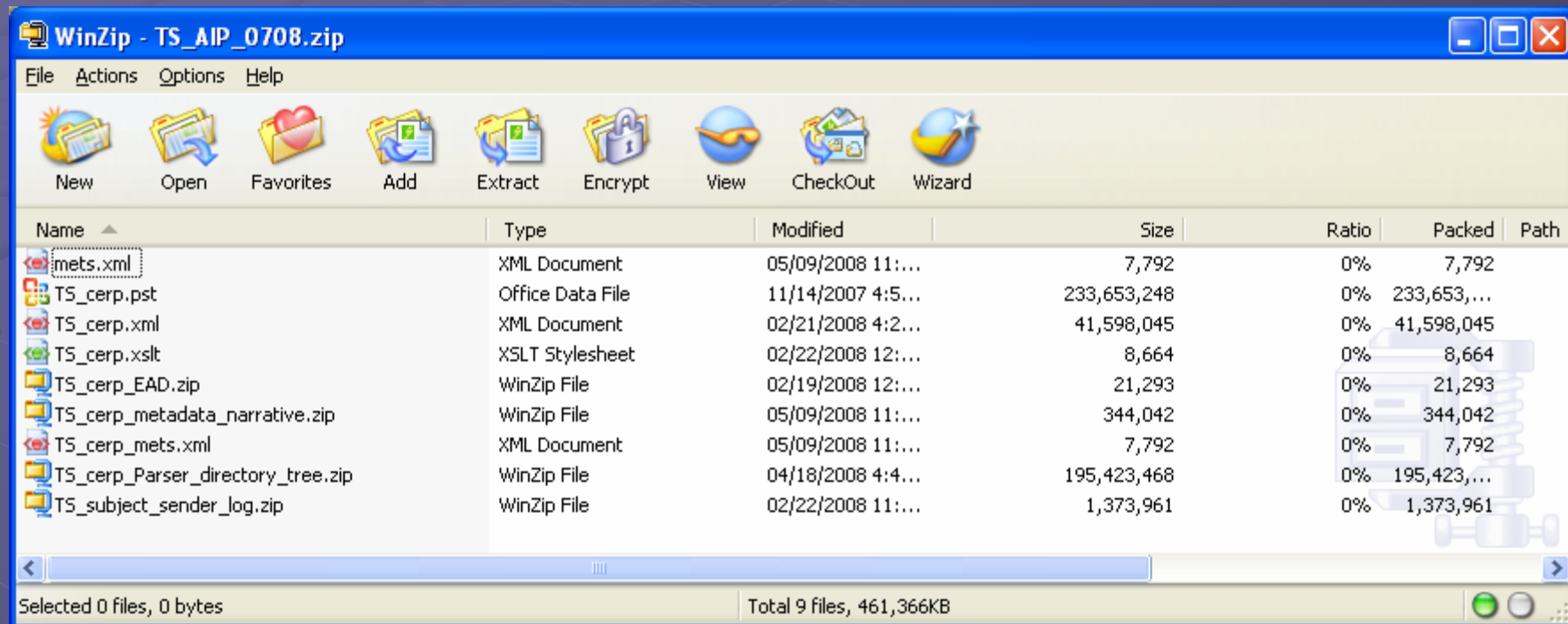
```
</MultiBody>
- <SingleBody>
  <ContentType>application/x-wordperfect6</ContentType>
  <TransferEncoding>base64</TransferEncoding>
  <Disposition>attachment</Disposition>
  <DispositionFileName>TestbedAgreement.wpd</DispositionFileName>
- <ExtBodyContent>
  <RelPath>./attach1790797246.xml</RelPath>
  <LocalId>1790797246</LocalId>
  <XMLWrapped>true</XMLWrapped>
</ExtBodyContent>
</SingleBody>
</MultiBody>
<Eol>CRLF</Eol>
- <Hash>
  <Value>6E81DB4AD3E8C8C5741087201905DD4405100D14</Value>
  <Function>SHA1</Function>
</Hash>
</Message>
```

Parser Subject-Sender Log

From	To	Date	Subject
"Ferrante, Riccardo" <FerranteR@CERP-PROJECT@SI-LISTSERV.SI>		Tue, 12 Dec 2006 10:22:05 -0500	[CERP] ACCEPTED: SAA 2007 Poster S
"Norine Goodnough" <goodnon@nancyadgent.com>	"Nancy Adgent" <nadgent@mail.rockefeller.edu>	Mon, 5 Jun 2006 11:00:15 -0400	FW: Poster
"Ferrante, Riccardo" <FerranteR@CERP-PROJECT@SI-LISTSERV.SI>		Thu, 22 Jun 2006 07:24:28 -0400	[CERP] Brief of presentation to American
"Ken Rose" <rosek@mail.rockefeller.edu>	"Nancy Adgent" <nadgent@rockefeller.edu>	Tue, 28 Mar 2006 14:43:07 -0500	revised Testbed Agreement
Nancy Adgent <nadgent@rockefeller.edu>	<Darwin Stapleton>, Ken Rose	Thu, 29 Jun 2006 14:50:00 -0400	Accession Documentation Forms
"Nancy Adgent" <nadgent@mail.rockefeller.edu>	<SchmitzfuhrigL@si.edu>, "Darwin Stapleton"	Thu, 29 Jun 2006 14:50:56 -0400	Accession Documentation Forms
Nancy Adgent <nadgent@rockefeller.edu>	<rossner@mail.rockefeller.edu>	Wed, 08 Mar 2006 17:23:00 -0400	Altered Images
"Norine Goodnough" <goodnon@nancyadgent.com>	"Nancy Adgent" <nadgent@rockefeller.edu>	Tue, 11 Apr 2006 09:48:28 -0400	brochure
"SAA Registrations" <registrations@dc2006meeting.org>	"SAA Registrations" <registrations@dc2006meeting.org>	Wed, 14 Jun 2006 13:30:04 -0500	PENDING: DC 2006 Joint Annual Meeting
"Nancy Adgent" <nadgent@mail.rockefeller.edu>	<SchmitzfuhrigL@si.edu>, "Darwin Stapleton"	Mon, 5 Jun 2006 09:10:02 -0400	Colloquium Photos
Darwin Stapleton <stapled@mail.rockefeller.edu>	<varianr@Rockefeller.edu>	Wed, 30 Aug 2006 15:53:19 -0400	Adobe Professional
Nancy Adgent <nadgent@rockefeller.edu>		Thu, 27 Jul 2006 13:58:00 -0400	Brochures for SAA
Steve Burbeck <sburbeck@mindspring.com>	Nancy Adgent <nadgent@mail.rockefeller.edu>	Thu, 19 Oct 2006 17:13:39 -0400	P.S. on attachment decoding
"Mark Conrad" <mark.conrad@nara.gov>	<elr@lists.archivists.org>	Fri, 23 Jun 2006 16:52:47 -0400	[elr] Annual Meeting of the Electronic Rec

MessageID	Hash	Errors	First Error Msg
<7B7FDF0E44197442A53FA1E3436963FA030DF9AE@SI-EDU>	FF5B99CE6D9E45B1406997C0E5FFB88975A58F9A		
<200606051500.k55F0JWd017935@smtp2.rockefeller.edu>	001EE7D33C18C56C561990131D33F325ECCC30FC		
<7B7FDF0E44197442A53FA1E3436963FAC20049@SI-EDU>	FC57017FD629C40132C6A9E06F71DAA4A3F81785		
<200603281941.k2SJf5qK004452@smtp1.rockefeller.edu>	6E81DB4AD3E8C8C5741087201905DD4405100D14		
	879185174 ED44F2B8CD80EEBAA06013240B59382EAC9B98E2		
<200606291850.k5TlopZ5013095@smtp2.rockefeller.edu>	FE013F16BCC45D2753E3FBFA54334B01191C4623		
	777908467 0BA76EB5B6AF25AEB6D3BACB1DF4976D7793B18A		
<200604111348.k3BDmXjs000633@smtp2.rockefeller.edu>	5B237DDEFC36E74C98EF262306C3E8083FBE1DC7		
<20060614-13300492-1848-0@fs2.webitacts.com>	ED7EEEF73C893B99B45AFF9582E9477BA1C3406		
<200606051310.k55DA0aA006099@smtp1.rockefeller.edu>	8749D07B0B9324E92834628B2C5297983D03C192		
<7.0.1.0.2.20060830155230.0326b3f0@mail.rockefeller.edu>	0D208E32351E7CD979CCF37B20BFAEB0A327843F		
	1866977147 A82D200F6C16C3EDD77CD3DD3ACE6BEB3398901A		
<4537EA83.7070306@mindspring.com>	C7AF4109623E5A1DFDAB240C7146E72438050155		
<s49c1c6b.004@smtp.nara.gov>	B6071152A54B6786917DFC844C243F762AF6293D		

Archival Information Package





Metadata Encoding and Transmission Standard (METS)

- Digital Library Federation initiative
- Non–proprietary standard
- XML format
- Supports many metadata schema
- Sections for descriptive metadata, administrative metadata, file groups, file hierarchies, and behavior



Completed METS AIP Form

AIP METS FOR DSPACE FORM <dmdSec>

Archival Term/DSpace Term/<Dublin Core Tag>

Collection/Community Name <dc:publisher>: Rockefeller Archive Center

Record Group/Sub-Community <dc:relation.ispartofpublisher>: CERP

Series/Collection <dc:relation.ispartofpublisher>: CERP Demo Mail

E-Mail Account Holder's Name <dc:creator>: Nancy Adgent

Department <dc:contributor>: Collaborative Electronic Records Project

Accession Number <dc:identifier>: 2008-CERP 1

Accession/Bundle Name <dc:title>: Nancy's Demo Set

Date Range, inclusive <dc:date>: 2006-03-08 to 2006-12-12

E-Mail Collection Folder Name(s) <dc:description.tableofcontents>: Nancy's In E
Rec Project

Subject Headings <dc:subject>: Electronic records
Email
Digital records
Projects
Smithsonian Institution Archives
Rockefeller Archive Center

CERP's DSpace



Smithsonian Institution Archives

THE ROCKEFELLER ARCHIVE CENTER

SEARCH

Go

[Advanced Search](#)

- [Home](#)
- [Browse All](#)
- [Browse Titles](#)
- [Browse Authors](#)
- [Browse By Subjects](#)
- [Browse By Date](#)
- [My Account](#)
authorized users
- [About DSpace](#)
- [Help](#)

[The Rockefeller Archive Center](#) >

Communities and Collections

Shown below is a list of communities and the collections and sub-communities within them. Click on a name to view that community or collection home page.

[Smithsonian Institution Archives](#)

[CERP Testbed 1](#)

[Email Records Testbed 1](#)

[CERP Testbed 2](#)

[Email Records Testbed 2](#)

[CERP Testbed 3](#)

[Email Testbed 3](#)

[Office of the Director](#)

[Email Records](#)

[The Rockefeller Archive Center](#)

[CERP Testbed](#)

[CD 1](#)

[CD 2](#)