Capturing the E-Tiger: New Tools for Email Preservation

Collaborative Electronic Records Project

Society of American Archivists
2008 Annual Meeting

THE ROCKEFELLER ARCHIVE CENTER

Smithsonian Institution Archives

The Collaborative Electronic Records Project (CERP)

- Design a preservation system and tools capable of preserving and maintaining digital records
  - A strong emphasis on email records
- Implement the system and tools at the partner organizations
- Produce a practicable preservation system model for use by other small to medium archives.

CERP Partners

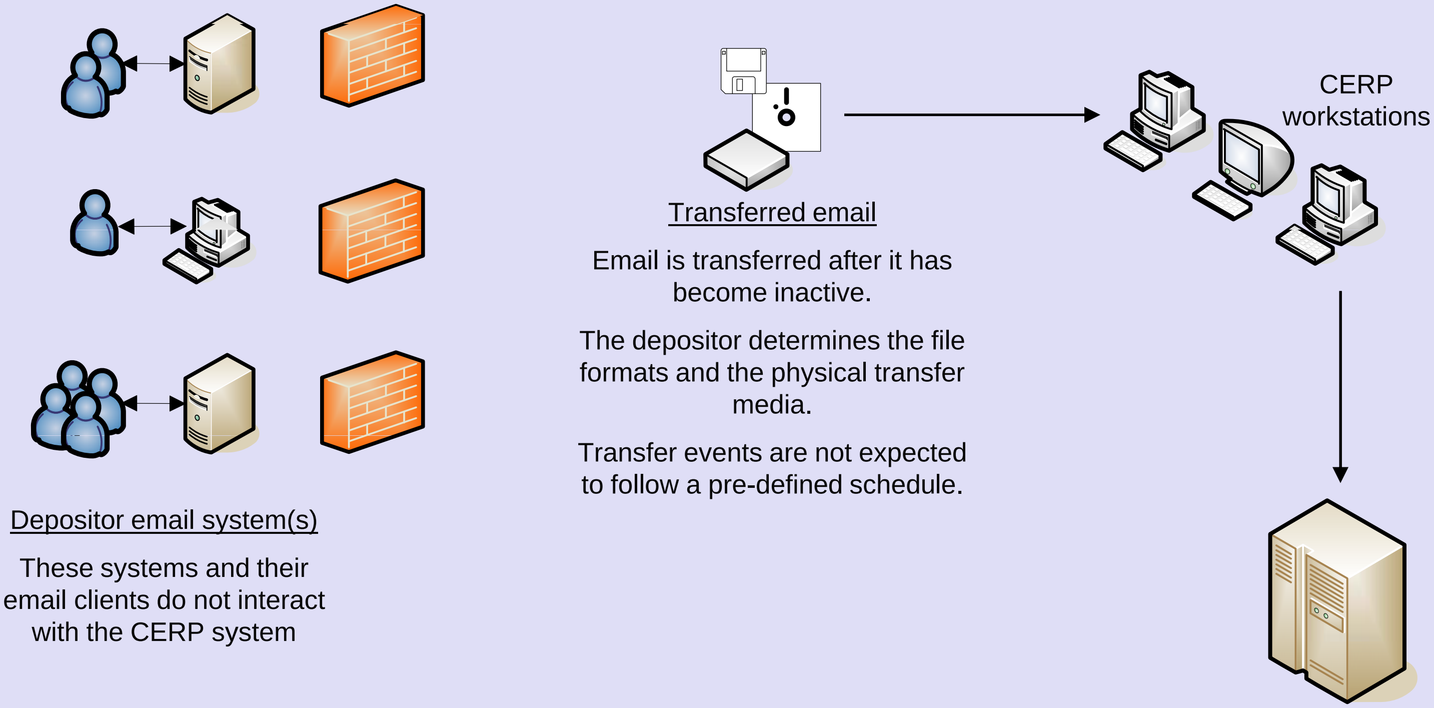
Rockefeller Archive Center

- Depositors include the Rockefeller family, their philanthropic and educational organizations, and non-family philanthropies.
- Little to no access to the depositors’ systems creating email or other digital records.

Smithsonian Institution Archives

- Depositors include the Institution, related persons and organizations, and other donors related to the history of American science
- Email transferred from a variety of systems, typically 5 years or more after becoming inactive
- Active digital preservation and curation program

Architecture



Depositor email system(s)

These systems and their email clients do not interact with the CERP system

Transferred email

Email is transferred after it has become inactive.

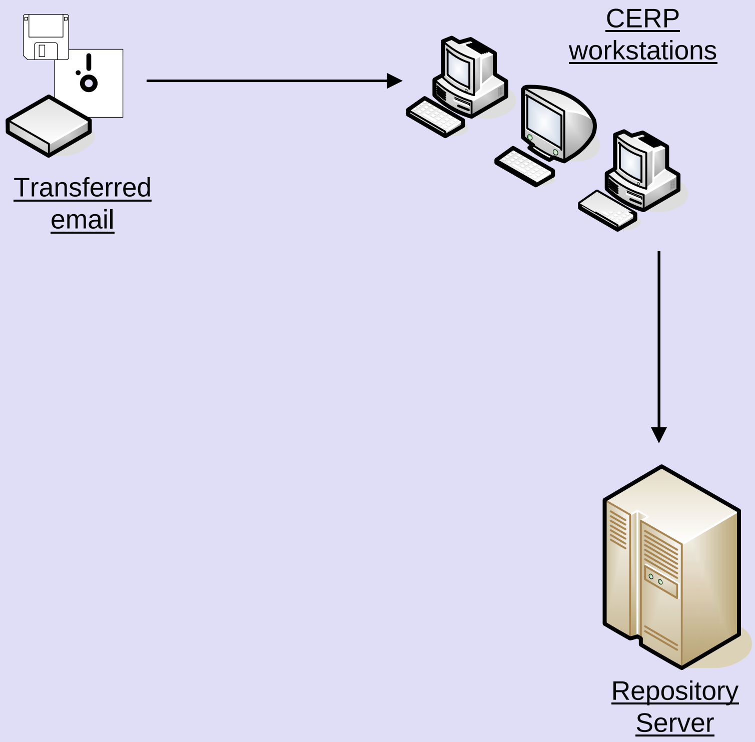
The depositor determines the file formats and the physical transfer media.

Transfer events are not expected to follow a pre-defined schedule.

CERP workstations

Repository Server

Architecture



If necessary, transferred email goes through a preliminary transformation into ‘mbox’ format.

An XML file of the email account is generated by the CERP Parser.

The XML is incorporated into the Archival Information Package (AIP) along with updated metadata information and Preservation Description Information (PDI).

The AIP is loaded into the Repository Server.

Choosing An Account Model

- Given a starting point of email messages selected by an account owner for archival deposit, relationships between those emails as well as any supplemental meaning that the owner has assigned through his/her organization of that account are valuable information that must be captured.
  - With a ‘message’ model, thorough documentation of each message, its interrelationships, and its context within the account is overwhelming in the face of email volume.
  - With an ‘account’ model, many of the relationships between the emails are already documented within the emails themselves. Further relationships, especially those assigned by the account owner, are present in the account structure and organization at the point of transfer.

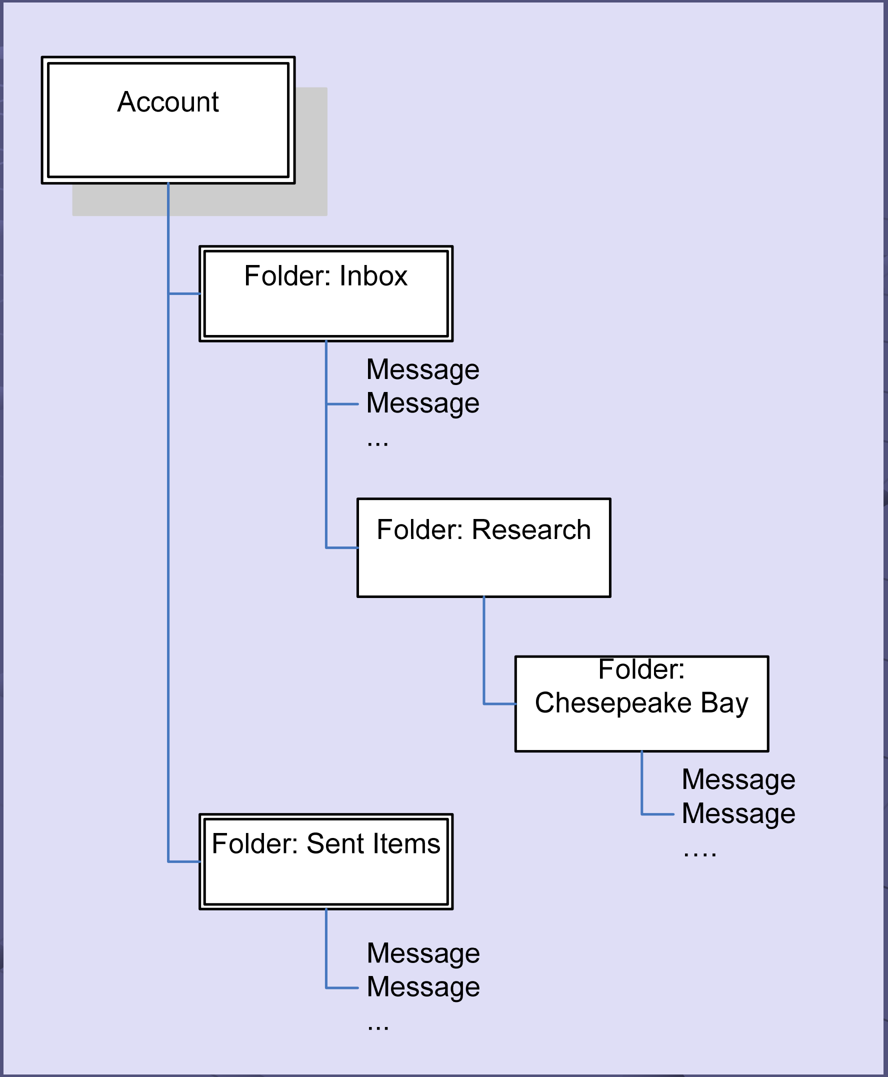
Email Account Preservation File

- Viability and risks of the native email format
  - Which one? How well is it documented? How long will software exist to read it? Which companies (if any) have a real commitment to stability and longevity?

- Choosing e**X**tensible **M**arkup **L**anguage (XML)?
  - XML is open, human readable and “self describing”
  - A good descriptive schema supports validity checking
  - There are many open source tools to create, manipulate and read XML

The Value of the Email Account Preservation (EMAP) Schema

- PRESERVATION:
  - A Schema defines how the XML tags for the various parts of an email relate to each other.
  - It is the Rosetta stone that guides how raw email is converted to XML



The Value of the EMAP Schema

- STORAGE:
  - Authorization filter to verify that an object purporting to be an authentic preserved email account is what it claims to be.
- SEARCHING:
  - Structure for subsequent search, display
  - Level of tagging enables deep data-mining
  - Cross-account searching, and possibly broader federated searches

From Transfer to AIP

- Various transfer methods
- Metadata gathering
- Attachment diagnosis
- Preliminary format transformation
- Final preservation transformation
- METS generation and final metadata
- AIP assembly

Using METS in the AIP

- Multiple types of metadata = excellent wrapper
  - DMDSec
    - Accession metadata stored in Dublin Core
    - Not limited to one descriptive metadata syntax
  - FileGroup
  - FileSec
  - StructMap
- Other information options available
  - AdminSec
- METS format for DSpace ingest

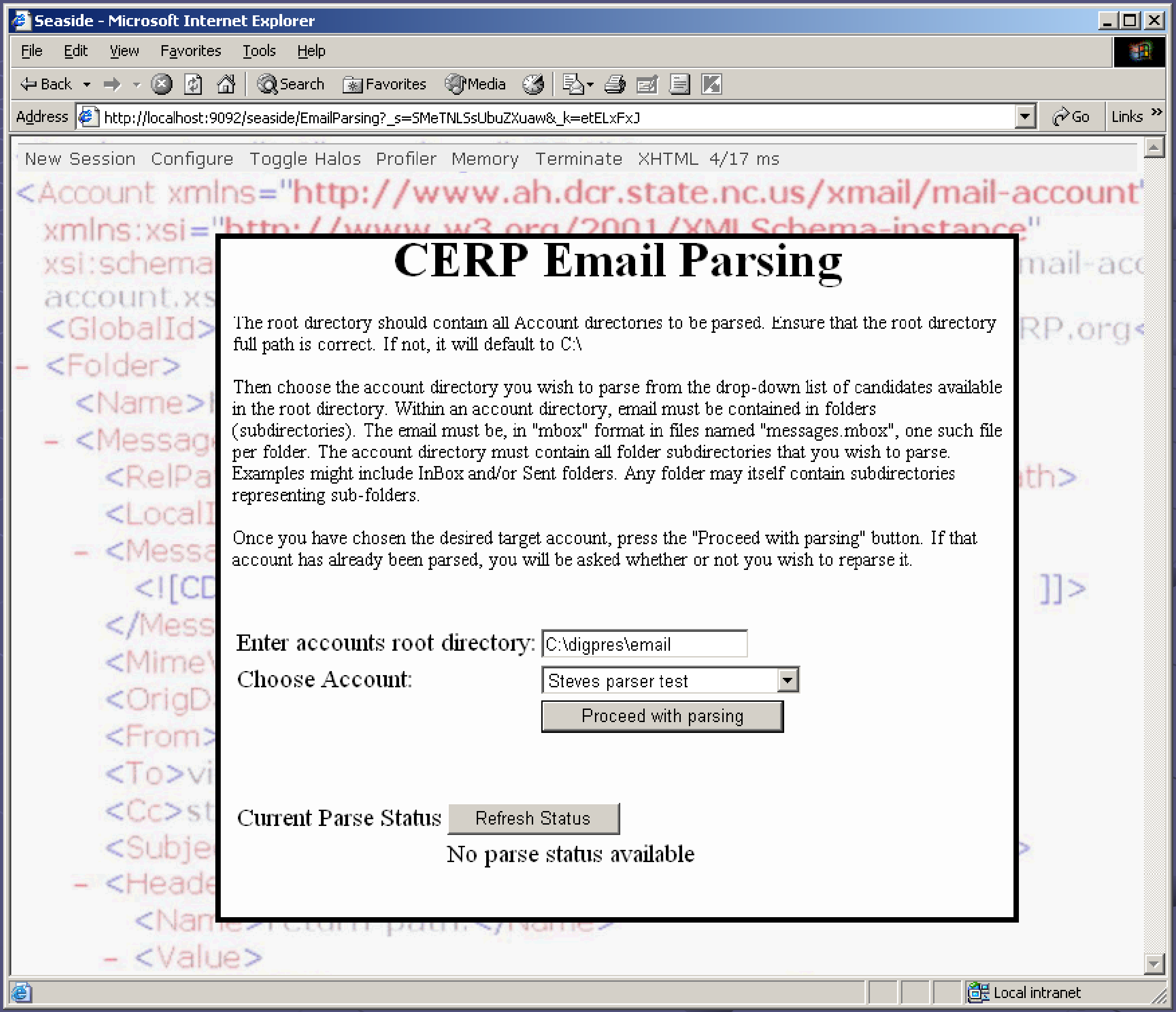
Email Conversion Results

- We have converted and validated 70 thousand messages in three test sets 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)
- Parse speed for an account with attachments
  - about a quarter gigabyte per hour on a Thinkpad T40 (March, 2008)

Variety is the Spice of Email

- Dozens of common email systems and 100s of others
  - We have encountered mail from Eudora (multiple versions), Simeon for MacPPC, Outlook/Exchange (multiple versions), AppleMail, Lotus Notes, Groupwise, Mozilla/Firefox, Pegasus Mail, and various Internet mail services such as gmail, Hotmail, YahooMail, Juno, and AOL   Each has its peculiarities.
- Some use non-standard date formats
- European and Asian mail may contain non-ASCII (actually, non UTF-8) characters
- Older email may have HTML in inappropriate places
- Forwarded and other “child” messages may be included in nonstandard forms

The Parser


Seaside - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address
Go Links
New Session Configure Toggle Halos Profiler Memory Terminate XHTML 4/17 ms
Local intranet

http://localhost:9092/seaside/EmailParsing?_s=SMeTNLSsUbuZXuaw&_k=etELxFxJ

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: C:\digpres\email

Choose Account: Steves parser test

Proceed with parsing

Current Parse Status Refresh Status

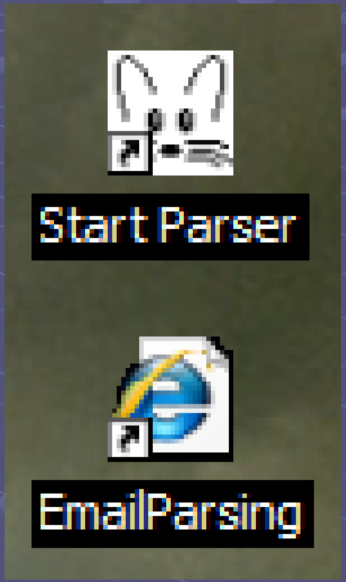
No parse status available

The CERP Email Parser

The Web Application Interface

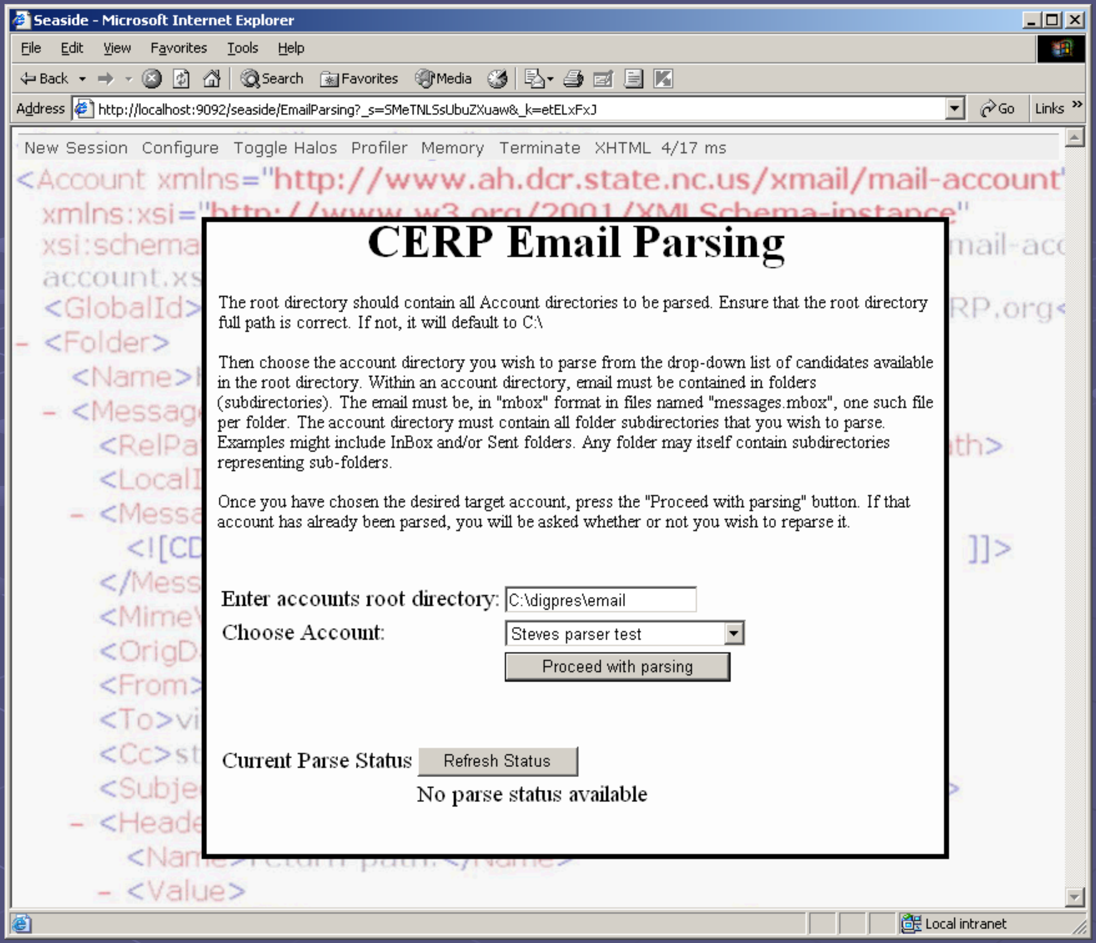
- The parser can be run from within Squeak, but most users will prefer to run it from a Web browser
  - The Web interface is built with a popular Squeak Web Application development framework called Seaside (www.seaside.st)
  - Seaside uses a web server (Comanche) that is embedded in Squeak.
  - Comanche is confined to supporting the parser and the Seaside application interface.

Running the CERP Email Parser



- Start the parser

- Start the Web UI
  - If necessary, start Seaside by executing “WAKom startOn: 9092”
  - The Web UI runs at http://localhost:9092/seaside/EmailParsing


Seaside - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://localhost:9092/seaside/EmailParsing?_s=SMeTNLSsUbuZXuaw&_k=etELxFxJ
Go Links
New Session Configure Toggle Halos Profiler Memory Terminate XHTML 4/17 ms
<Account xmlns="http://www.ah.dcr.state.nc.us/xmail/mail-account"
Local intranet

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: C:\digpres\email

Choose Account: Steves parser test

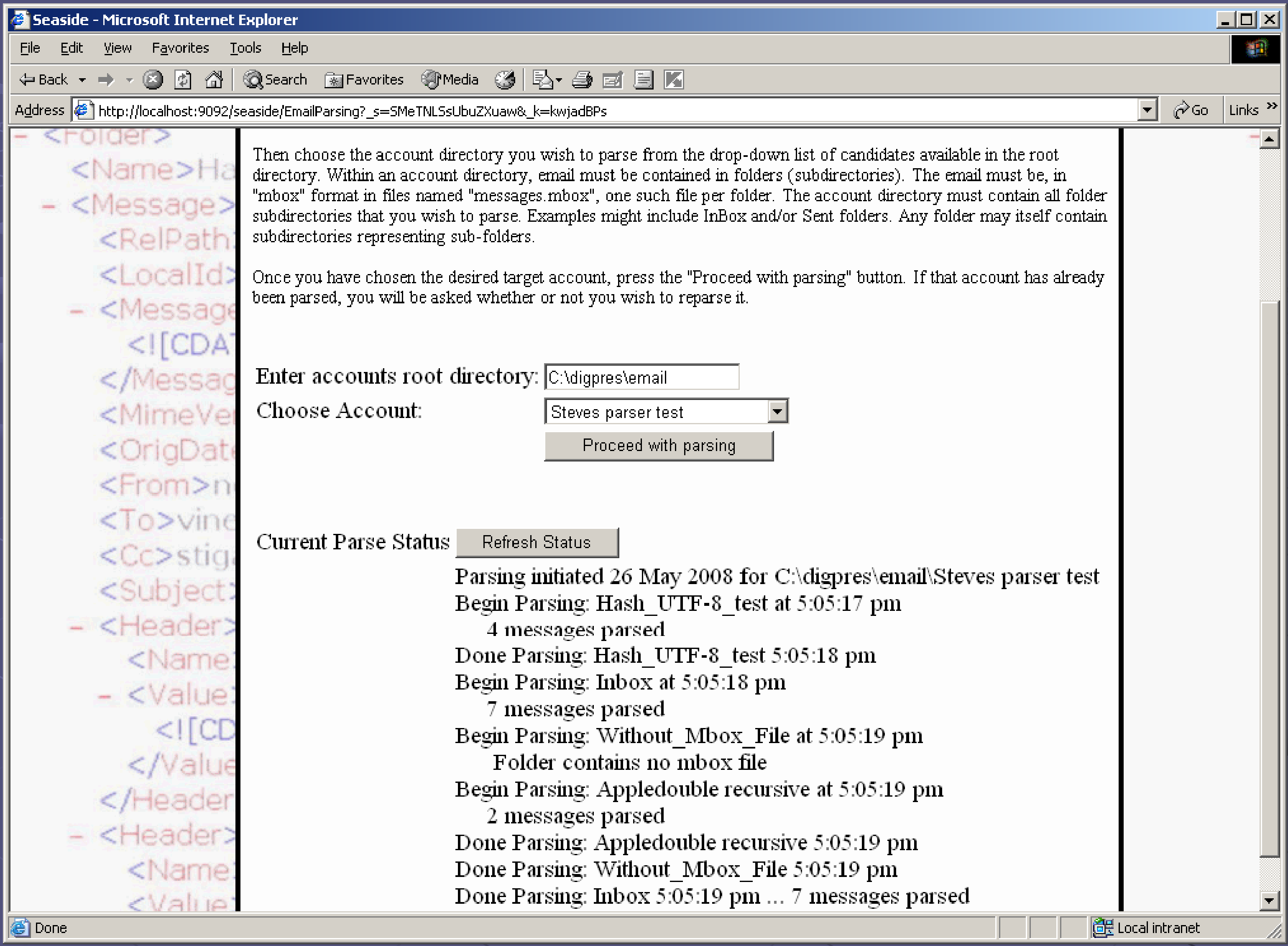
Proceed with parsing

Current Parse Status Refresh Status

No parse status available

- Navigate to the directory containing the prepped account
- Select the account folder
- “Proceed with parsing”

Parsing Results Status



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: C:\digpres\email

Choose Account: Steves parser test

Proceed with parsing

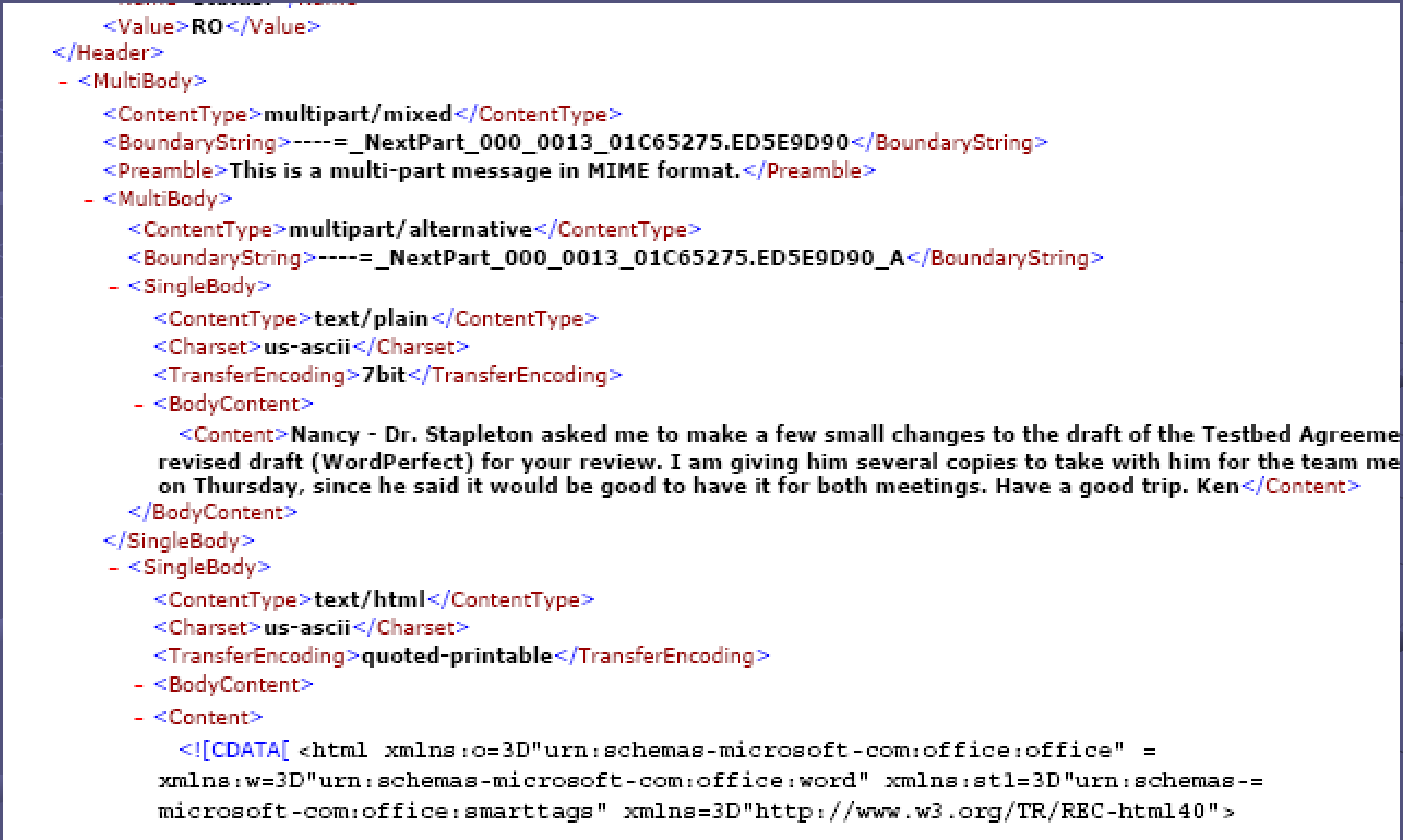
Current Parse Status Refresh Status

```
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
```

Preservation AIP

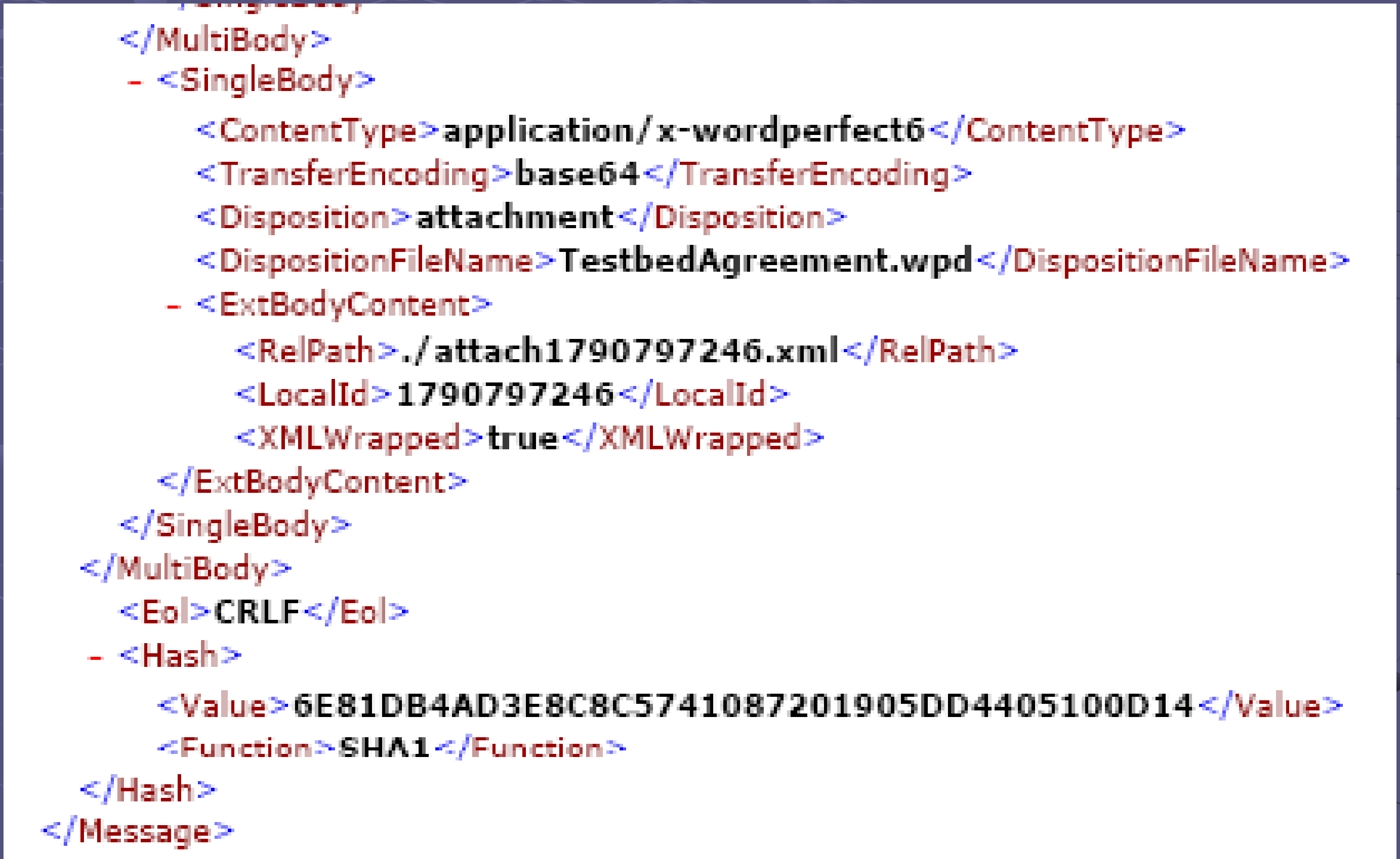
- Source File(s)
- Accession Metadata
- Preservation Description Information (PDI)
- Preservation File(s)
- METS File

Parsed E-mail Body Excerpt



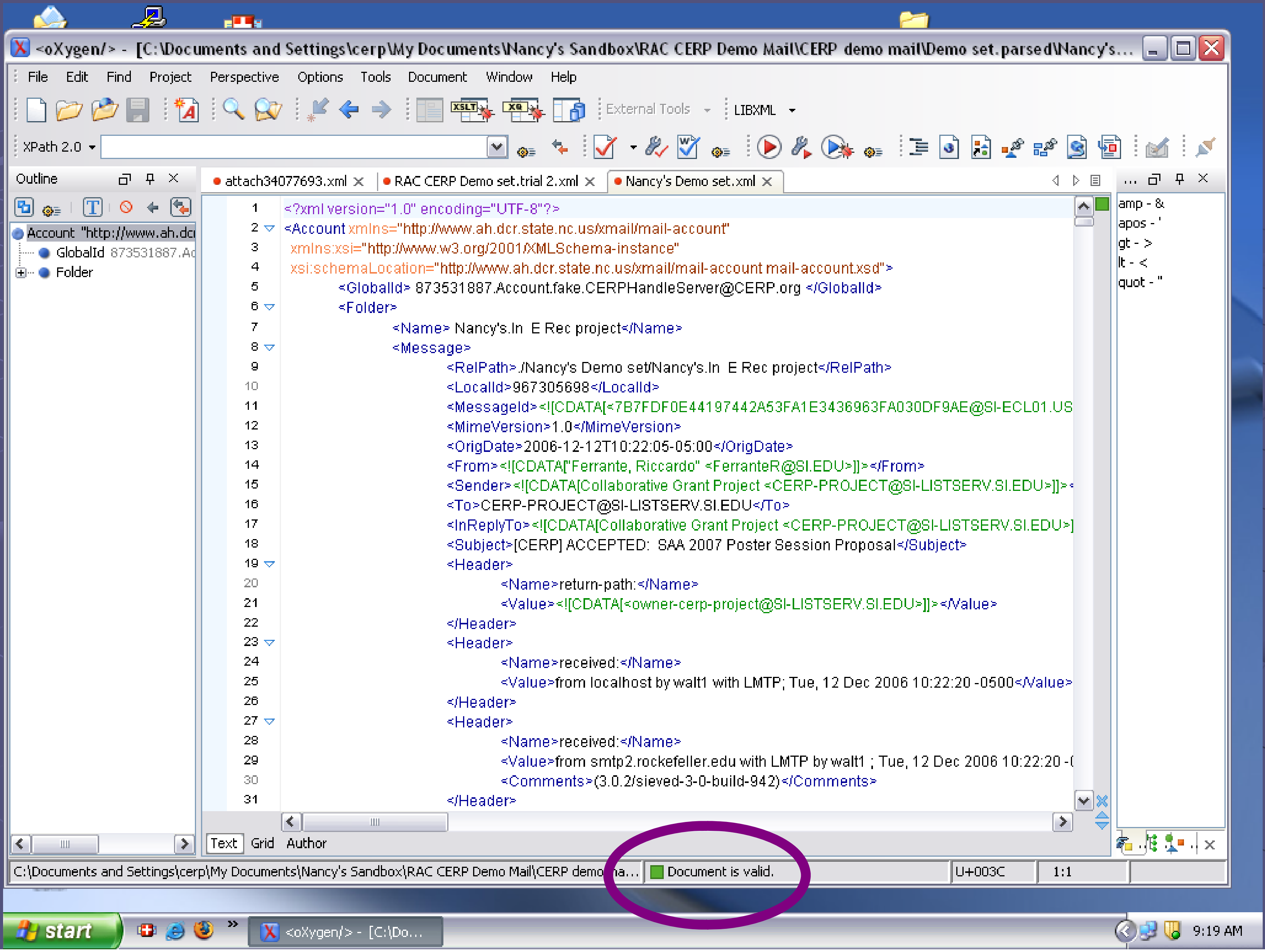
```
      <Name>Status</Name>
      <Value>RO</Value>
  </Header>
 - <MultiBody>
      <ContentType>multipart/mixed</ContentType>
      <BoundaryString>----=_NextPart_000_0013_01C65275.ED5E9D90</BoundaryString>
      <Preamble>This is a multi-part message in MIME format.</Preamble>
   - <MultiBody>
        <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 Agreeme
          revised draft (WordPerfect) for your review. I am giving him several copies to take with him for the team me
          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



```
      </SingleBody>
    </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>
```

Validation Message



<oXygen/> - [C:\Documents and Settings\cerp\My Documents\Nancy's Sandbox\RAC CERP Demo Mail\CERP demo mail\Demo set.parsed\Nancy's...

```
<?xml version="1.0" encoding="UTF-8"?>
<Account xmlns="http://www.ah.dcr.state.nc.us/xmail/mail-account"
 xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
 xsi:schemaLocation="http://www.ah.dcr.state.nc.us/xmail/mail-account mail-account.xsd">
        <GlobalId> 873531887.Account.fake.CERPHandleServer@CERP.org </GlobalId>
        <Folder>
                <Name> Nancy's.In  E Rec project</Name>
                <Message>
                        <RelPath>./Nancy's Demo set/Nancy's.In  E Rec project</RelPath>
                        <LocalId>967305698</LocalId>
                        <MessageId><![CDATA[<7B7FDF0E44197442A53FA1E3436963FA030DF9AE@SI-ECL01.US
                        <MimeVersion>1.0</MimeVersion>
                        <OrigDate>2006-12-12T10:22:05-05:00</OrigDate>
                        <From><![CDATA["Ferrante, Riccardo" <FerranteR@SI.EDU>]]></From>
                        <Sender><![CDATA[Collaborative Grant Project <CERP-PROJECT@SI-LISTSERV.SI.EDU>]]><
                        <To>CERP-PROJECT@SI-LISTSERV.SI.EDU</To>
                        <InReplyTo><![CDATA[Collaborative Grant Project <CERP-PROJECT@SI-LISTSERV.SI.EDU>]
                        <Subject>[CERP] ACCEPTED:  SAA 2007 Poster Session Proposal</Subject>
                        <Header>
                                <Name>return-path:</Name>
                                <Value><![CDATA[<owner-cerp-project@SI-LISTSERV.SI.EDU>]]></Value>
                        </Header>
                        <Header>
                                <Name>received:</Name>
                                <Value>from localhost by walt1 with LMTP; Tue, 12 Dec 2006 10:22:20 -0500</Value>
                        </Header>
                        <Header>
                                <Name>received:</Name>
                                <Value>from smtp2.rockefeller.edu with LMTP by walt1 ; Tue, 12 Dec 2006 10:22:20 -(
                                <Comments>(3.0.2/sieved-3-0-build-942)</Comments>
                        </Header>
```

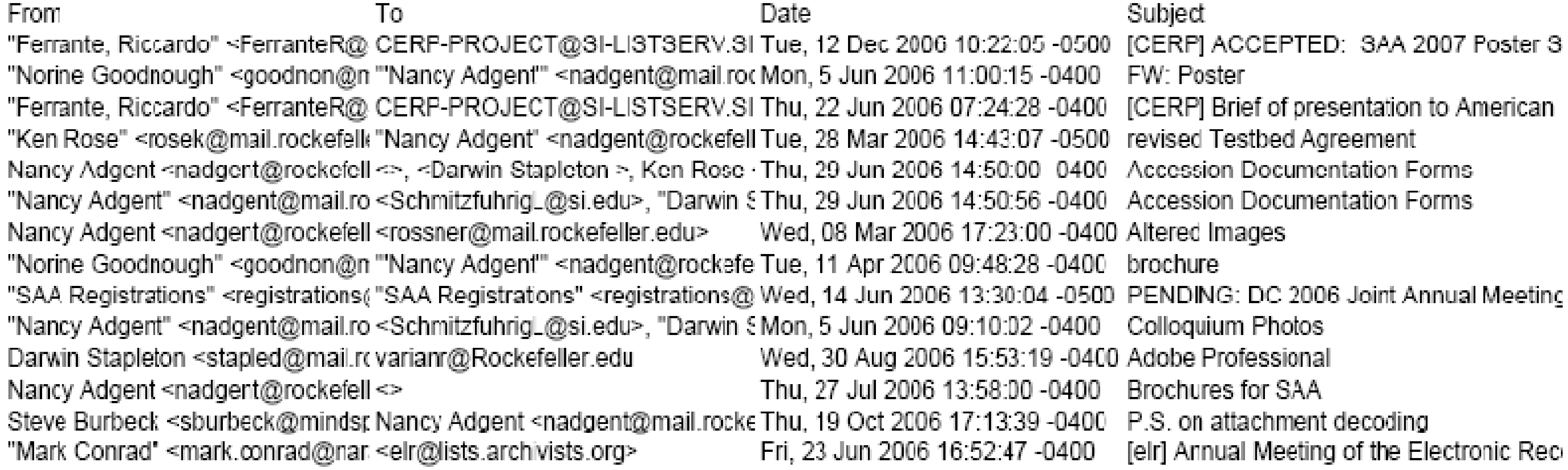
C:\Documents and Settings\cerp\My Documents\Nancy's Sandbox\RAC CERP Demo Mail\CERP demo ma...

🟩 Document is valid.

U+003C

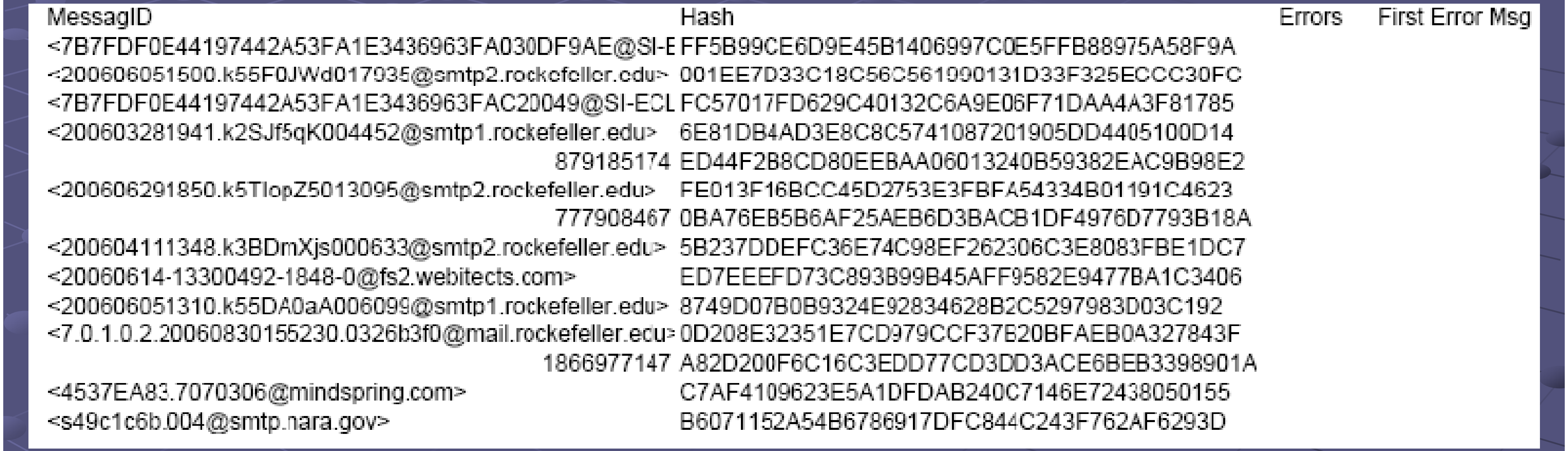
1:1

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@n	"'Nancy Adgent'" <nadgent@mail.roc	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.rockefell	"Nancy Adgent" <nadgent@rockefell	Tue, 28 Mar 2006 14:43:07 -0500	revised Testbed Agreement
Nancy Adgent <nadgent@rockefell	<>, <Darwin Stapleton >, Ken Rose <	Thu, 29 Jun 2006 14:50:00 0400	Accession Documentation Forms
"Nancy Adgent" <nadgent@mail.ro	<SchmitzfuhrigL@si.edu>, "Darwin S	Thu, 29 Jun 2006 14:50:56 -0400	Accession Documentation Forms
Nancy Adgent <nadgent@rockefell	<rossner@mail.rockefeller.edu>	Wed, 08 Mar 2006 17:23:00 -0400	Altered Images
"Norine Goodnough" <goodnon@n	"'Nancy Adgent'" <nadgent@rockefe	Tue, 11 Apr 2006 09:48:28 -0400	brochure
"SAA Registrations" <registrations(	"SAA Registratons" <registrations@	Wed, 14 Jun 2006 13:30:04 -0500	PENDING: DC 2006 Joint Annual Meeting
"Nancy Adgent" <nadgent@mail.ro	<SchmitzfuhrigL@si.edu>, "Darwin S	Mon, 5 Jun 2006 09:10:02 -0400	Colloquium Photos
Darwin Stapleton <stapled@mail.rc	varianr@Rockefeller.edu	Wed, 30 Aug 2006 15:53:19 -0400	Adobe Professional
Nancy Adgent <nadgent@rockefell	<>	Thu, 27 Jul 2006 13:58:00 -0400	Brochures for SAA
Steve Burbeck <sburbeck@mindsp	Nancy Adgent <nadgent@mail.rocke	Thu, 19 Oct 2006 17:13:39 -0400	P.S. on attachment decoding
"Mark Conrad" <mark.conrad@nar	<elr@lists.archivists.org>	Fri, 23 Jun 2006 16:52:47 -0400	[elr] Annual Meeting of the Electronic Rec

Parser Subject-Sender Log (cont.)



MessagID	Hash	Errors	First Error Msg
<7B7FDF0E44197442A53FA1E3436963FA030DF9AE@SI-E	FF5B99CE6D9E45B1406997C0E5FFB88975A58F9A		
<200606051500.k55F0JWd017935@smtp2.rockefeller.edu>	001EE7D33C18C56C561990131D33F325ECCC30FC		
<7B7FDF0E44197442A53FA1E3436963FAC20049@SI-ECL	FC57017FD629C40132C6A9E06F71DAA4A3F81785		
<200603281941.k2SJf5qK004452@smtp1.rockefeller.edu>	6E81DB4AD3E8C8C5741087201905DD4405100D14		
879185174	ED44F2B8CD80EEBAA06013240B59382EAC9B98E2		
<200606291850.k5TIopZ5013095@smtp2.rockefeller.edu>	FE013F16BCC45D2753E3FBFA54334B01191C4623		
777908467	0BA76EB5B6AF25AEB6D3BACB1DF4976D7793B18A		
<200604111348.k3BDmXjs000633@smtp2.rockefeller.edu>	5B237DDEFC36E74C98EF262306C3E8083FBE1DC7		
<20060614-13300492-1848-0@fs2.webitects.com>	ED7EEEFD73C893B99B45AFF9582E9477BA1C3406		
<200606051310.k55DA0aA006099@smtp1.rockefeller.edu>	8749D07B0B9324E92834628B2C5297983D03C192		
<7.0.1.0.2.20060830155230.0326b3f0@mail.rockefeller.ecu>	0D208E32351E7CD979CCF37E20BFAEB0A327843F		
1866977147	A82D200F6C16C3EDD77CD3DD3ACE6BEB3398901A		
<4537EA83.7070306@mindspring.com>	C7AF4109623E5A1DFDAB240C7146E72438050155		
<s49c1c6b.004@smtp.nara.gov>	B6071152A54B6786917DFC844C243F762AF6293D		

Long-term storage – Using DSpace

- Selected for expediency
- Significant limitations
- Surmounting the scale and access obstacles will require further research
- Other DSpace projects may generate some solutions

Preservation Issues

- Complex account structures
- Hierarchical structures
- More than just email formats
- Email standards and adherence
- Email system idiosyncrasies

Loose Email “Standards”

- RFC2822 and other standards are a good start that handle most cases.
- Yet email continues to evolve and standards continue to lag.
- To be widely adopted, lagging standards must support virtually all preexisting practices…an impossible goal without compromises that are open to interpretation.
- Different email client vendors interpret the standards differently.
- And there are the inevitable mismatches between interpretations (and inevitable bugs).

Preservation Lessons Learned

- 100% success is an unrealistic goal
  - Some emails are just too broken to parse without manual intervention
- We ***can*** achieve at least 99.9% success (and save the few unparsed emails for human inspection)
  - This error rate is not unlike physical archives
- The EMAP Schema provides a very robust structure that can support sophisticated and complex access and retrieval

Next Steps

- Continued testing
- Review by others
  - Parser and documentation on CERP website
  - Considering ‘webinar’ events
- Testing with email-related records
  - e.g., mailing lists
- Identifying/developing search tools
- Integrating privacy/sensitive data solutions

http://siarchives.si.edu/cerp

Rockefeller Archive Center

Nancy Adgent, Project Archivist
NAdgent@rockarch.org
914-366-6355

Smithsonian Institution Archives

Riccardo Ferrante, Project Manager
FerranteR@si.edu 202-633-5906

Lynda Schmitz Fuhrig, Project Archivist
SchmitzFuhrigL@si.edu 202-633-5917