

## Preservation of Electronic Mail

Druscie Simpson

The primary goal of this proposed project was to research and test methods to create a facility that would allow us to preserve email independent of the software used to create the record. The North Carolina State Archives currently possesses six gigabytes of email from Governor Jim Hunt's administration. This email has been appraised and culled to create the core email for preservation. However, the email originally existed as a Microsoft Outlook .pst file and could only be accessed using the version of Microsoft Outlook supported by Microsoft. As a result of this project we tested various programs that are in the development stage to convert email from Microsoft's proprietary .pst format to a non-proprietary and stable XML format.

More specifically the project aimed to test the implementation of a server-base facility using the IMAP protocol to collect e-mail messages to be converted. Most modern email clients can be configured to not only receive mail from the user's primary server, but to also have an additional IMAP server configured so that selected email can be sent to a secondary store. We needed only to develop the components that would be responsible for taking the contents of this secondary store and exporting it as a set of XML files.

A committee was formed that was made up of various members of the State Archives staff to assist in this project. It consisted of Druscie Simpson, head of the Archives Information Technology Branch; David Minor, applications programmer; Kelly Eubank, electronic records archivist; Mark Valsame, governor's papers archivist; and Paul Kiel, temporary fellowship programmer. This committee met once a month to evaluate progress made and to exchange research results and further ideas on dealing with the preservation of electronic mail.

A great deal of progress was made on the "Preservation of Electronic Mail" project during the funding period. The design of the overall system was completed, including the XML Schema for the core XML content files and well as the code responsible for converting e-mail messages to XML. Attachments are not yet handled but we have a clear plan for adding this functionality. During the final quarter of the grant period we accomplished the following tasks:

1. Completed the core software conversion "engine".
2. Built a simple and secure control center to allow the administration and selection of accounts to archive, and to report on the status of each account.
3. Testing
4. Documenting the software and associated processes.

During the course of our work, we quickly discovered that the XMTP open source project was an inadequate starting point for the following reasons:

1. The XML it produced was too "open". No DTD or XML Schema could be created that would validate all documents without using the "any" construct.
2. The attachments were stored within the XML file. Our goal is to manage these attachments as native files. The attachments can be accessed and managed better after they have been extracted from the e-mail messages.
3. The XMTP open source project does not have support for e-mail messages conformant to older e-mail standards.

In addition, the XMTP open source project had virtually no activity since its introduction in July 2001.

Timeline of activity and results:

Activity	Notes
Set up budget accounts for grant	This was more involved than expected, but has been accomplished. The Dept. of Cultural Resources is receiving grant monies into accounts for travel, personnel and benefits, and equipment.
Complete arrangement and description of Gov. Hunt's email	The appraisal and arrangement of Governor James B. Hunt's office email has been completed and a finding aid is completed.
Prepare email on server and create master and working copies to be exported to XML	Completed
Recruit and hire programmer on contractual basis to develop, test and create documentation for software components	Programmer hired February 1.
Purchase low-end Windows 203 server and Visual Studio.net 2003.	Completed
Install, configure and become familiar with the operation of the following softwares: 1. hmailserver (open source) 2. Apache web server (open source) 3. php (open source) 2. Visual Studio 2003.	Completed
Create index/finding aids of email to be used for access purposes	Completed
Analyzed the implementation strategy and interface design of the xntp java library, in preparation to create a .net version of the same	The xntp java library was found to be very simplistic. Instead we used the standard .net IO and XML libraries and wrote the parsing code from scratch.

Become familiar with the object model exposed by the hcom Com component that is part of the hserver installation.	We are successfully using hmail server to discover and access each e-mail message for each account.
Demonstrate that the hmail server has been setup and that it is receiving messages via IMAP.	Done.
Determine whether or not the hcom component can supply the necessary information to reconstruct the original e-mail message, including all attachments, in its entirety. If no, then create a new modified version of hcom that can.	The hmail server stores messages in RFC2822 format as individual files on the file system. The hcom component is responsible for the registration of those files to each account. It is a simple matter to use standard file IO to access the messages. No modification to hcom or the mail server is necessary.
Develop a version of the xmtmp component in C#. Determine whether to use the SAX or DOM model for XML processing.	An XML Schema has been completed, and the basic message parsing and conversion to XML has been completed.
Create a prototype program that will take each message in a specified account on the hmailserver and then by using hcom to access the original message, build an XML version of that message with the xmtmp component.	The software to convert and save each attachment and the related file has been developed. The software to manage the "housekeeping" data related to account status has been developed.

Next steps:

We still need to complete the method by which we will handle attachments. Also, the documentation needs to be completed. We then would like to install the application on the individual computers in the Archives and Records Section to fully test the application against real time emails. Once that testing is completed and further tweaking of the software is done, we would like to present our solution to the Information Technology Services (ITS) agency for state government to get their feedback. Hopefully we will be able to have ITS adopt our application and incorporated it into the North Carolina government ncmal system.