
Progress Report: March 1, 2007- August 31, 2007

NHPRC EMCAP Tool

Kelly Eubank



NORTH CAROLINA DEPARTMENT OF CULTURAL RESOURCES

MICHAEL F. EASLEY, GOVERNOR
LISBETH C. EVANS, SECRETARY

OFFICE OF ARCHIVES AND HISTORY
JEFFREY J. CROW, DEPUTY SECRETARY

TO: Lucy Barber, National Historical Publications and Records Commission
FROM: Kelly Eubank, Project Director, Electronic Records Archivist

SUBJECT: North Carolina, Kentucky, and Pennsylvania's Electronic Mail Preservation
Collaboration Initiative, Grant Number NAR RE-05701-07

Progress Report: March 1, 2007-August 31, 2007

The Electronic Mail Preservation Collaboration Initiative aims to build a prototype tool for the collection and preservation of archival electronic mail. Staff from the states of North Carolina, Kentucky, and Pennsylvania have teamed together to leverage their knowledge and skill sets to help design and test a tool that will collect electronic mail, maintain a copy of the original email and the attachment, as well as transform the email itself from its native file to an XML file based on an XML schema. In addition, once the tool has been designed and tested, we will deploy this tool to participating pilot partners for phase two testing and feedback. The project team includes: Kelly Eubank (NC), project director, Druscilla Simpson (NC), head, IT Branch; David Minor (NC), programmer; Chris Black (NC), archivist; Ed Southern (NC), head, Government Records Branch; Mark Myers (KY), electronic records archivist; Glen McAninch (KY), head Technology Analysis and Support Branch; Linda Avetta (PA), information technology generalist administrator; Cynthia Bendroth (PA), head Division of Records Administration and Imaging.

The tool, when fully developed, to the user will not appear dramatically different from current client configuration. The account for the Email Collection and Preservation (EMCAP) will appear as an "Archives Folder" next to local folders and incoming mail server. The user, when ready to archive the email, will simply drag and drop that folder or email to folders under the "Archives" folder or directly into the Archives folder. The only difference with the Archives folder is that the Archives folder and all sub-folders will be configured to go to the EMCAP server which will be located at each respective state archives' archival repository. The collaborating partners will develop training material to explain how to use this tool.

MAILING ADDRESS
4610 MAIL SERVICE CENTER
RALEIGH, NC 27699-4610

TELEPHONE 919-807-7280
FAX 919-733-8807

LOCATION
109 EAST JONES STREET
RALEIGH, NC 27601

Objective 1: Information Technology Environments

1. Continue the development of the e-mail collection and preservation software to convert e-mail from its native format into the more stable XML format and complete written documentation for the program.

Information Technology Environments:

The project faced some interesting challenges from the beginning. The staff in North Carolina had not administered a multi-state grant. As such, we had several meetings with people in the budget office and the Chief Financial Officer to clarify how the grant would be administered and how reimbursements could be issued to employees from other states. Once those issues were resolved, the team set about planning the opening meeting and its agenda.

On March 21, 2007, members from the partner states met in North Carolina to discuss the information technology environments within their respective states, how the proposed tool would work in those environments, to further define the tool's capabilities, and to discuss each state's roles and responsibilities with the grant (see the Meeting Minutes in Addendum A).

Pennsylvania:

Both Pennsylvania and Kentucky have very centralized IT environments that use Microsoft Office Suite and Microsoft Exchange (Outlook 2002). Any information technology project has to involve and receive approval from the state central IT agency. The staff in Pennsylvania has been talking with that group and preparing documentation to share with them. The Technical Architecture Review Board has to approve any IT project. Of particular concern is the fact that it is a beta system that will be used in a production environment. The IT staff in Pennsylvania suggested setting up a separate exchange email server. The email, when copied, would be copied to this intermediary server and could then be sent to the Email Collection and Preservation (EMCAP) server. The primary concern from IT is the use of open source software and allowing beta to connect to production which poses risks on a production environment.

Kentucky

Kentucky also has very centralized IT. Every agency in Kentucky is on a Microsoft Exchange Server system that is managed by Commonwealth Office of Technology. Microsoft is the approved architectural standard. The staff in Kentucky has tested the idea of incorporating a database from the previous governor's administration, which contains the text of emails from constituents, as well as information on the resolution of the issue raised. The current governor also used this mail log system for the first few years of his administration. David Minor, the programmer in NC, has been working with the staff in Kentucky regarding a correspondence database that Kentucky received from the previous governor, to try to see if the correspondence could be reconfigured as .pst files and dumped into the hmail IMAP server too. After analysis of the data, it was concluded that the e-mail record could not be effectively imported into HMailServer, because the information was too incomplete to fully reconstruct the e-mail record. However, the constituent mail database, which also contains scans of incoming paper correspondence (in tiff format), will need to be converted to a less proprietary format that is consistent with the grant's database structure. This procedure is outside the main purpose of the grant. Customized mail logs such as this will likely be a challenge for archivists in the future.

Kentucky is proceeding with plans to import official correspondence in Microsoft Office personal folders (.pst) format into the HMailServer environment. We set up an account in the Kentucky Department of Libraries and Archives (KDLA) Commissioner's Office to test the regular transfer of records via an existing account. These files will not be imported until a full archival appraisal of the .pst files has occurred and all non-archival files are removed. Kentucky also plans to determine the HMailServer networking configuration that is most compatible with the state IT architecture should the opportunity arise in the future to implement the EMCAP tool on accounts that are not under KDLA's control.

North Carolina

North Carolina has both a decentralized and centralized IT environment. The executive level departments (those departments in which the cabinet secretary is appointed by the governor) use the services of the statewide Information Technology Services (ITS). They are not required to do so and can in fact implement their own email system, but many of them do outsource these services to ITS. The Constitutional offices, those with elected cabinet level offices, are not bound by rule to ITS services. A small number of them chose to use the ITS email system; others chose their own email system. The service offered by ITS has a web mail component but employees can and often do use client software to access and send email. The staff in North Carolina desire to test the EMCAP with Microsoft products, GroupWise, and possibly Lotus Notes.

Technical Requirements

Using grant money, both North Carolina and Pennsylvania received servers in order to host the EMCAP service. Still outstanding is how to set up the instance in Pennsylvania. David Minor, programmer in NC, researched the suggestion of using an intermediary server between the production Exchange server and the EMCAP server. His conclusion is that Exchange does not support two accounts for one person very well.

Three scenarios may be feasible:

- A. The user "pushes" mail to a secondary account on a separate server IMAP server, EMCAP pulls from those secondary accounts.
- B. The user "pushes" mail to a secondary set of "shared" folders on the same e-mail system. Each user is given their own shared folder for this use. EMCAP pulls from those secondary folders.
- C. The user subscribes to a service that allows the user to choose the folders on the existing server from which mail should be archived. A custom piece of software installed on each user's computer periodically "pulls" new mail from those folders and pushes it to the EMCAP server.

Scenario A was the original plan, and this is what is addressed by the Invitation For Bid (IFB). This solution will work for every mail environment.

Scenario B is the alternate scenario that has gotten the most attention lately. This solution requires a significant amount of programming for each different server platform--Exchange, Domino, and GroupWise.

Scenario C is the alternate scenario that was discussed at North Carolina's first face-to-face meeting with Pennsylvania and Kentucky. This solution requires a significant amount of programming for each client platform: Outlook, Notes, and the GroupWise Client.

Scenario C was found to be unworkable because it is very difficult to insure that the software:

- a. Did not include any malware agents, either currently or in any future release.
- b. Would not impact the performance of any user's workstation at any time.
- c. Would only copy mail from certain designated folders.

Scenario C, however, requires the least amount of work on the user; in fact individuals who already organize their mail into folders would not need to make any changes, and those folks who were not accustomed to using folders would only have to begin moving important mail into a single pre-defined folder. Once the security, trust, and performance issues have been successfully resolved, this solution will be just as easy for the IT administrative group as Scenario B. Scenario A requires the most demands on the IT administrative group, due to the fact that each user has to have their e-mail client configured with an additional account.

At this time, we are still working through the issues with Pennsylvania and exploring alternative methods.

Objective 2: Potential Partners

2. Test the software to determine its scalability, efficiency, and employee compliance, using records retention schedules.

Potential Partners

Pennsylvania:

Pennsylvania would very much like to work with two entities for project—the Governor's Office of Correspondence and Pennsylvania's Historical and Museum Commission (PHMC) Executive Office, of which Barbara Franco is the executive director. The internal partners, the internal executives of PHMC have agreed to participate. At this time, the question still outstanding with regards to beta software being used in a production environment need to be resolved.

Kentucky:

The staff in Kentucky has the email from the outgoing commissioner of the Kentucky Department of Libraries and Archives (KDLA) as well as the agreement of the current commissioner of the KDLA to participate. They also plan to work with the Secretary of Education Cabinet and her legal staff regarding a disk of .pst files left by her predecessor.

North Carolina:

The staff in North Carolina met with the Deputy Secretary of State Haley Haynes and the Chief Information Officer of the Secretary of State's office, Bruce Garner, in September to discuss more in depth the tool and that office's possible participation (see Addendum B—Consultation Summary). Both Ms. Haynes and Mr. Garner were keenly interested and agreed to participate. This work will dovetail nicely with work they are currently doing to update their retention schedules. Ms. Haynes suggested the Corporations Office and possibly the Notary Office as possible candidates for

participating. North Carolina identified the April/May timeline for their involvement. In addition, the staff in NC has identified a potential office within the Governor's office as partners. They have not formally agreed to participate in the project; they have expressed an interest in participating and have been provided with an overview of the project and technical specifications. Future meetings are scheduled to discuss it in detail more with that office. North Carolina is also considering approaching the Statewide Emergency Management (SEMA) office. The State Agency and University Records Unit of NCSA has been working closely with this office to schedule its records. Staff from SEMA has inquired about automating some of the transfers.

Also during this time, staff in North Carolina performed a basic count of email messages that we currently receive from the Office of the Governor, Community and Citizen Services (via manual transfer sneaker net). North Carolina hopes to compare the number of constituent emails transferred to the Archives via sneaker net to those that may be transferred through the automated process.

Of note there was a marked difference between the folder systems in the two most recent transfers of email to the Archives. The first transfer contained a folder system that relied on numerous subject terms (40 separate folders at the second level; the first level was simply 2006) and in some instances was as many as five folders deep. The most recent transfer contained only 3 separate folders: Agency Responses, No Response Needed (NRN), and Sent Items. When asked if the office had transferred everything it intended to, they said that they had and had moved to a "big bucket" approach and were using searching tools to locate specific emails within each of the three large buckets.

Presentations

At the Government Records Section meeting of the annual meeting of the Society of American Archivists on August 31, 2007, Mark Myers, Electronic Records Archivist for KDLA, did a short presentation regarding the project. The presentation, put together by Kelly Eubank with input from David Minor, gave a short introduction to the project and its goals, offered screen shots of what the tool may look like as well as diagrams of the configuration, and introduced the launch of the project Web site (see below.) Myers also discussed the issues the project has faced so far (the IT environments in KY and PA) and the challenges those issues have presented. There was lively discussion, centering on the issues of letting end-users decide what is to be kept, and the presentation was well received.

Web site Launch

Chris Black, Archivist in the Electronic Records Unit, worked diligently this cycle to put together a Web site of the project, <http://www.ah.dcr.state.nc.us/records/emailpreservation/>

The Web site currently contains a summary of the project and its goals; contact information for key personnel; minutes, diagrams, other related materials from the kickoff meeting; and hyperlinks to similar email preservation initiatives. An event time line will be added in September.

Objective 4-5

4. Test ways of providing access to these XML files and the feasibility of doing so, whether through existing online catalogs, a third party vendor, or web interfaces.

5. Extract and save attachments in their original, as received, format as a native stand-alone file, which will be wrapped in XML. The association between message and attachment will be kept and will allow navigation from message to attachment back to message.

Programmer Update

At the outset of the grant, the staff in North Carolina, charged with further development of the tool, intended to hire a full time programmer for one year to complete the programming. Staff posted the job announcement on known IT project websites and in the Career Placement offices and Computer Science departments at local universities and community colleges. Unfortunately, we did not receive applications from qualified applicants in this area. Given the time line, the staff decided it needed to change direction in order to get the project completed. We explored the idea of doing the project through a contractual services agreement. The IT staff of the North Carolina State Archives and Records Section put together the design specifications for the tool. After doing research, the staff decided to submit the specifications through NC's online purchasing tool to ITS. ITS has an employee services program that has a number of programmers registered who are looking for work with the state. We will begin this phase October 1, 2007. In taking this approach, the programmer will not have to do as much of the documentation, and, therefore, to prepare the work should take a much shorter time period to complete--4-6 months.

Possible Microsoft Testing

Given the centralized IT environments in Pennsylvania and Kentucky, the partners did discuss the possibility of working with the new Microsoft Office 2007 tools and the Microsoft Exchange 2007 environment. In this environment, users can set rules for email and documents to automatically "journal" the files. The files are sent to "managed folders" on the Exchange server. However, given the time constraints and the difficulties in getting buy in and hiring a programmer, the partners decided it was not feasible to test this capability during the grant period.

Addendum A

NHPRC Email Project

Kickoff meeting

March 21, 2007

Attendees: David Minor, Chris Black, Druscie Simpson, Kelly Eubank, Linda Avetta, Cindy Bendroth, Mark Myers, Glen McAninch, Ed Southern

Introductions:

Kelly Eubank: NC, Electronic Records Archivist, Electronic Records Unit. Responsible for state and local agencies

Chris Black: NC, Archivist, Electronic Records Unit, Responsible for records management (governor's office), and processing of state and local records

Linda Avetta: PA, Information Technology Generalist Administrator. Responsible for electronic records including email, no policy right now (rescinded to be put into a different format) and is currently being reviewed by attorneys, digital projects for web, enterprise management systems, Filenet/Omnirim integration for electronic records (OmniRIM currently manages physical records at the State Records Center and all records schedules), working with OGC to write policy; for several years have asked for an electronic archives system but not funded; working with another agency and VideoBank (national geographic files) about using their system for storage and management of electronic records, video and audio files.

Cindy Bendroth: PA, Appraisal and Accessioning Section, Head. Responsible for appraisal and processing; email policy; state agencies acquisitions

Glen McAninch: KY, Technology Analysis and Support Branch, description and management of electronic records, staff of 4 now reduced to 3; Persistent Archives Project in San Diego 3 years; electronic records archives for state pubs and minutes and governor's records snapshots (speeches, press releases, photos, etc.) on the web; Governor's Office is transitioning to a content management system that was developed to handle congressional [constituent?] correspondence, Governor's website was contracted out (NIC) along with a single portal and search engine, developed for 8-10 states. However, the Governor's Office appointed their own webmaster about a year ago.

Mark Myers: KY, Responsible for training on electronic records for state and local agencies; Currently working on a general schedule on correspondence--"up to 2 years" vs. "retain for 2 years" so in a big education push now regarding email, general correspondence has 2 year retention, official correspondence has a permanent retention,

- Awareness that retention now in the hands of the individual employee more than before.
- KY has an open records law: if an agency denies a request the requester can appeal to Attorney General; what is personal? Anything on a state computer is a public record. KDLA is pulling things out that are not business and putting them on their general schedule under the series "non-business related correspondence".
- "Electronic messages" series on the General Schedule for Electronic and Related Records lets them discuss email as temporary, electronic messages vs. correspondence. Allows more flexibility with retention. Includes voice messages, text messages, blackberries, helps avoid "partial messages" that refers to emails sent, even though email has been deleted, etc. also "peer to peer" messages.

- May use former state librarian's email or the new cabinet secretary's email.

David Minor: NC, applications programmer. Responsible for server and network support, SQL software for online catalog, imaging support, METS metadata for preservation

What is going on in states?

NCSA:

Regularly receives email (every 6 months) from the Office of Citizens and Community Affairs in the Gov. Office, primarily consisting of constituent correspondence

- our email is centralized as a courier IMAP service;
- we have collected the previous Superintendent's email from the Department of Public Instruction (DPI). DPI uses GroupWise to manage their email;
- Kelly is the chairperson of NCMail users group, the email system offered through Information Technology Services (ITS) which wants to get away from being a "docstore". ITS is proposing that they store all email for 7 years and then delete everything. Kelly is trying to educate ITS about records retention. ITS is only interested in discovery requests and wants to save everything for ease. Calendaring didn't work the way folks wanted it so NCMail is looking at using Outlook for calendaring (\$7.00 per person per month). NC doesn't have a centralized active directory. ITS offers the use of Novell, which has very few users; agencies tend to use active directory within each agency without intermingling active directories.

Kentucky:

- KDLA is currently receiving high definition and low definition audio and video files in 8 mb/seconds video format.
- GIS (raster database 1 tb before they went to color) working with them to accession records. Creative Services Department heavily uses Sony (Blue ray). They transferred to Archives professionally created DVD's that are encrypted. KET (Kentucky Educational Television) also talking about an ERM system. KT is digitizing for access all their film holdings of on-the-air programs for preservation because their film is deteriorating. At the moment, the archives is storing files on server to provide access in a media player format.
- Email is centralized through IT in KY and everyone uses Exchange servers.
- KDLA encourages agencies to use the auto archive feature of Outlook for records that are retained 2 years or less. Beginning to centralize servers at State IT as well; also centralizing some IT staff. Past governor was the first to be able to serve 2 terms and therefore is able to have his policies in place for 8 years. The current governor is changing things, including issues of centralized IT.
- KDLA's email guidelines were worked into the state's IT architecture standards. If a server fails, their ITS agency replaces it. Email boxes are limited on the email server, and the default storage space is the user's C drive although many agencies store personal folders on network drives. A lot of Mark's email training is really more about how to use Outlook for management than about email itself.

- KY is giving a big training push for agencies to take full advantage of foldering in Outlook to ease some problems associated with the retention of emails. They are showing examples of what is archival, general correspondence, etc. KDLA helped staff in the president of university identify what was archival and what was not. Yes it goes here, no it goes there sort of thing.

Pennsylvania:

- In 2008, PA is upgrading their hardware and software to move directly from Exchange 2000 to Exchange 2007.
- Currently utilizes a single active directory with multiple domains, with some exceptions. PA also faces the problem that if you change jobs within state government, your email goes with you even though it belongs with original job.
- Email boxes are spread across many servers; workstations are set up differently across agencies and even differently within some agencies - some default to the C drive, some don't. Many users whose .pst folders are configured on the C drive don't understand that when their pc crashes their email is gone, unless they back it up.

Discussion of Implementation of hmail collection:

David: suggested placing the collection server placed with the agency, where you have control over the Exchange server. Where you don't, put server at ITS. Security is with secure socket connections (using encryption), which is very easy to do...you don't have to write anything. Firewall and filtering depends on the agency. The agency can do it on their own or contract with ITS. Interagency is done individually on an agency-by-agency basis.

Glen: Email is stored as an XML file in Hmail server? Not Outlook file? KDLA uses a DMZ and Microsoft Active Directory, agency by agency, which are then linked together. Where should the email server sit? Inside or outside firewall and DMZ? KDLA likes the fact that their web server is inside the DMZ.

David If everyone you want to collect email from is inside private network, then put it inside, and if anyone is outside, then put it outside. David is not sure how we would handle providing access via the web server.

Glen thought we would go that far but it actually will be the next grant. **David** says we should be able to use current search tools to "check out" email from the repository.

Overview of email system:

Repository based on the OAIS model; **David** not concerned with that except to be aware of the requirements.

Local server with Outlook client on their computer:

Person opens Outlook and email, sees what is on the mail server, what is on their pc, and what is on the hMail server. Mail can be stored on collection server instead of your local drive if you would rather.

In the client-centric model, the user creates folders based on series descriptions, moves or copies messages from server to the archival collection server via IMAP. Even on the collection server they can delete, etc. until the snapshot is taken. Every night (or at a predetermined time) we take a

snapshot of all folders on the hMail server. Once this has been done the email can't be deleted. A server-centric approach could also be used.

Server collection center

It will be placed between the email server and the hMail server. It has to be able to log onto the email server so will have to know each individual's password. **David** thinks hMail can use the active directory account, which would solve the password issue (KY and PA use active directory) but he needs to double check. Server-based APIs will be utilized.

- User would create a folder inside Exchange server called archive mail (with sub-folders). We would capture only the archive folder. It will not work to capture the auto archive folder, because we would have to write custom software to re open the .pst archived file and resave it.
- If you hold pre-existing .pst files like Kentucky, you have to take existing .pst files and reload them to Outlook and then send the email to the hMail server. Macros could be used to do this, so you wouldn't have to parse the .pst file.

Outlook 2007 gets away from the .pst files, but we will still have to deal with them for a while. **Glen** thinks it would be a good part of this project to go ahead and develop the scripts needed to handle .pst files, etc. and load them back into Outlook.

For users that refuse to be records managers, we could capture all their folders as if they were archive folders.

Incremental vs. Automatic Capture?

Transcript of archival process for accessions. When email is in the intermediate collection server, it is not accessioned. The intermediate collection server is basically a records center. Before we accession the email, we will submit a list of what we have captured and have the submitting agency sign off on it. The email can be changed or deleted until it has been archived. We can set up rules that say that we officially archive every X number of days (or hours) from the collection server. So if the user deletes any messages while they are still in the collection server, then those messages won't be captured.

David: North Carolina is trying to not do server-centric simply because we would need to deal with it agency by agency because every agency is different. Updates would have to be installed on every server, etc. We'd rather do client-centric and then if Pennsylvania could do server-centric, we could test both scenarios.

The least burden on the user, the better. The contract Programmer could write a script to create archive folders on the Exchange server automatically (would need more storage quota). Synchronizing prevents the collection of duplicate messages. The server-centric model would only require the use of one server.

The email is transformed to XML at the collection server level.

The intermediate storage sever will contain the original email file, the XML file for each account, and all the attachments. All attachments will be broken out of the email as primary artifacts, and will be brought into the repository in their native format.

Set up a directory for each domain. Database will give information about each account (starting date, etc.).

Discussion of Software Capabilities

Automatic metadata extraction:

Software only captures the header information. Background transactional information will not be automatically captured, but we will have places to store that information (address, starting and ending date, etc.) If an employee transfers to another position in the same domain, we would stop the old account, clean it up, and remove it from the email server. We can then create a new account within the domain and start capturing again.

Address book data capture:

Software does not capture address book data. The consumer could request that it be transferred along with the messages, but the automatic capture of address book data is beyond the scope of this project. Another option would be to have each employee submit all their .pst files when they leave their job, which would include their address book, as it was at the time they left.

Preservation of look and feel of email:

The look and feel of the email will not be preserved as part of this project. The features will be captured (font, background color, etc.), but you won't see them when you look at the email. *David* doesn't know what happens to the digital signature, but will look into it. 2003 allows digital signatures and encryptions, but these features are clunky. 2007 had encryption built in to click in automatically.

Digital rights management (DRM):

DRM is handled via the synchronization. We will have to find a way to unencrypt things that have been encrypted. If a message is set to disappear (which actually becomes encrypted) in 4 hours, it will have to be able to be unencrypted.

Mark wondered if you could possibly read a message, but not print or resave it somewhere else. If a message is not accepted by the hMail server because it is encrypted, etc., the user will receive a message that the transfer/copy didn't work and to try again in a different format. Restricting the use of DRM will be a matter of setting policy and a training issue. However, the recipient may not know that a sender has used DRM. HMail doesn't do anything to the attachments. HMail doesn't deal with any problems, issues of DRM because it just captures the email as it is. Really need policy that says any state agency should not accept any records/submissions with DRM attached. *Kelly* reported that Microsoft feels like DRM is really needed by industry, which doesn't have deal with the issues surrounding public records.

Unable to Read Email

If the email is not readable, contributor may get a warning message. If the attachment is unreadable, contributor may not get a warning. When you try to turn the message into XML, then you will definitely get the warning message. So the user needs to deactivate any DRM before they save to the archive store.

Need Adobe Acrobat at the server level to convert attachments to pdf/a for each state. For some of us that should be just an upgrade.

Set up of hMail servers:

Kentucky will depend upon their ITS, plus Glen. SQL server used more than MySQL.

Pennsylvania is not sure yet. Both are on Exchange server, so will only need one collection server. Can use MSDN license since it is just a test. Can also use the full version of SQL server. Also runs off of Windows XP, but hope to use Windows Server 2003. All will run on an XP or Vista box, but those probably won't have enough storage.

Encryption/Secure Sockets Layers:

Question: If email server is behind the firewall, is it important to have encryption? Partners following this scenario do not need Secure Sockets Layer (SSL).

- this scenario makes it easier. Kentucky has all Exchange servers in the same place; thinks their ITS will still want some kind of filtering, whether it is SSL or something else.
- SSL will not keep unwanted people out, but will show that there is something encrypted. Active directory does all the monitoring for you. Each partner will probably want SSL once we go beyond the pilot stage.

Email with access restrictions:

Software can capture email with access restrictions. Providing access to it will prove problematic. When we capture the email, a copy will remain on the user's desktop. If someone wanted to access that email the archives could refer patron to the agency to see the copy there. Once the email is in the repository, you can set up the repository to restrict certain files.

- *David* says we have a lot of opportunities for the user to place restrictions, type notes, etc. if need be through a web-based management process. Then the archivist could go through and see if the material is truly sensitive and why and for how long (e.g. construction negotiations.)
- It would be best if the archivist could apply restrictions at the folder/subfolder level, not at the message level

Training:

Mark would like to have all this set up (using Exchange server) before we start training. *Glen* reported that ITS is looking at some sort of records management application to handle email, but questions of what, who will pay for it, etc. have not yet been resolved. KY's ITS is only interested in the storage issue. It should be inviting to KY's ITS to have this solution available in order to, at least, deal with archival email

Kelly spoke about the tools developed by the Managing the Digital University Desktop (MDUD) project. Online tutorials are available on the website created by that project's team. The tutorial helps to identify what is a record, provides decision trees, etc. and how long the email record should be kept as a result of the identification. Visual rendering often gets the point across better.

Confidentiality restrictions may prove to be a major issue to be dealt with during the training component.

Linda: When you are in the broken out mode, is it still in the subfolder? Yes, so it would be a good place to apply the retention at that subfolder. *David* said another option would be to have the repository be a full-blown records management location, rather.

Mark: DRM. New computer had a trial version of Office 2007. If you used it, and didn't upgrade at the end of the trial version, it locked your documents as read only. If you purchase a copy of Office 2007, it will not unlock the files unless you have bought it online. Similar to the difference between word and word perfect.

What do we want to see in the program?

Hashing:

Kelly: Is there any kind of hashing as part of the program?

David says not yet, but it could be added to it as part of the process. **David** thought hashes were used at either end of a communications channel, but **Kelly** explained it is also being used for authentication purposes, in case the document's trustworthiness is ever challenged. So we would run the hash between the original on the original server and the original on the collection server and again on either end of the XML transformation.

David noted that the hashing needs to be stored elsewhere, so that the person who has access to the document doesn't have access to the hash.

- We essentially need 2 "safes," one for the documents and one for the hashes and have an infrastructure in place that doesn't allow access to both the documents and the hash along with audit logs. In fact, we really do not need access to the hashes at all except to save new ones. Secure sockets can be set up to do it intrinsically. **David** will create it as soon as it is feasible.

Q: When we create the XML message, we are dropping the email standard messages?

A: We will be preserving the original bit stream. Anytime left over after development will be devoted to improving the search functionality.

Ongoing Support:

Q: Linda: Since is it open source, which Pa's IT is uncomfortable with, who will maintain and update the product?

A: David...not just using open source, but also building on it.

Q: Glen: We are making it as portable as possible, following standards for databases.

A: David: You will have to change ITS' mind because you will periodically need programmers to keep it up to date. It would be possible to find a software development company to take over the development and support. We also plan on registering all our development with SourceForge. Need to contract with ITS from the very beginning, so that hopefully they will provide that support. We really haven't come to the point of having to find that ongoing support. HMail DOES have support though.

Q: Glen: Is anyone else working on this?

A: David says not that he is aware of. Whenever **David** talks to companies about it, and that we are talking about more than 10-year retention, they back off and say that is not what they are doing. It is the long-term retention that isn't being developed. If anyone needs an IT shop, it is archives because there isn't anything off the shelf that meets their unique needs.

Searching:

Q: Glen: What search functions will be available by the end of the grant?

A: David says not much, but whatever time he has left he will devote to that. You certainly can do basic searches like you do for current email when you pull the email accounts back into Outlook.

David: Once you get it into the repository we need to build a system to search the email. You will not be able to search across all accounts until we develop something to parse the XML. You could also push it to SharePoint and then you would be able to search across multiple accounts. **Linda:**

Since you can save the attachments as pdf's you will also be able to do keyword searching across all the attachments. **Glen:** .pdf's don't permit really elaborate searches.

Q: Linda: Often as part of discovery requests they want every single email with this topic or from this person, etc.

A: David that is easy to do once it is in XML. You can write a special purpose tool that you can use to answer those types of questions. Also, the native body of the email is saved as a text file (not XML) and the XML acts as a server that points to the text file. **Linda:** Vista can search across everything. **Glen** said he has a copy of Vista but has not tested the searching feature yet.

Kelly: It is like what we had to do with Archive-IT, if we had multiple collections you couldn't search across the collections. So we had to place all of our harvested websites in one collection.

Glen: It is really the difference between basic searches and intelligent searches?

David Some rudimentary browsing and searching would certainly be needed to help even manage the emails. The "owner" will want to be able to search and determine whether they have already put a particular email into the archive store or not. 10 years from now, when the discovery tools have changed, it will be easier to use whatever the attorney, judge, etc. wants to use.

Linda: Penn. is looking at a discovery tool for the state, which is challenging because there aren't email policies in place yet.

Q: Mark: Can this be expanded to include Blackberries?

A: David thinks so as long as it is being done through an Exchange server, but we don't have plans to include it at this time. Any peer-to-peer transfer wouldn't be captured. Since it is client based, it could be used for Hotmail or gmail, etc. We could write software that reads Hotmail automatically, but again not written into this grant.

Q: Linda: on the Email Preservation Project document, page 5, what does David mean about placing limits on the message store?

A: David is referring to the dwell time on the collection storage server. Penn. doesn't want users to have access to email once it is on the collection server, so their dwell time is zero. A copy will be on the Exchange server for as long as the user needs it and then will go directly to the repository server, so they will not need additional storage space.

Q: Linda: Virus protection. Penn. is using Postini so email doesn't even get to the inbox. Catches 98% of spam. Where is the virus protection occurring in the hMail project?

A: It is being run on the hMail server by hMail. It is configurable to use different providers. If there is a virus, it still captures the email. Why? It is possible that the email is still very important independently of the virus. In Penn. Postini captures spam email including those with viruses. There are 2 sections on Postini: non-virus spam, virus. User can choose to send messages from Postini to their Outlook inbox anyway. You see what the spam filter collected and make a decision on the message.

Mark: Filters in Ky. might mark something as spam because it is more than 3 people on the distribution. Folks consuming the email at the end of the process will have to run virus detection. HMail server will have to have up-to-date virus detection. **David** wonders if it is necessary on server if you already have good detection software in place before it gets to the hMail server.

David: one thing virus detection does is to strip out the attachment and run the detection on the attachment. We will be able to do it on our file store. We'll be looking at these issues as we go along in the grant.

Q: Linda: Does the 64 bit factor of Vista make a difference?

A: David thinks it shouldn't since email has to be interoperable. Will probably have more issues with Encryption than 64-bit.

Q: Linda: Under IMAP description...are the API's being developed?

A: Yes, **David** says they have been developed. You can add and delete accounts through the API's that are already done.

Possible partnerships in each state:

Kentucky:

- Planning on using the email of an outgoing commissioner and current commissioner of KDLA.
- Will also try to meet with the Secretary of Education Cabinet and her legal staff about a disk of .pst files that her predecessor left.
- Will also try to sell the current Secretary of Education Cabinet to participate. KDLA says if it is official correspondence it is frequently printed on paper, with a letterhead.
- KDLA has received the previous governor's files that were in a visual basic database that includes scanned images and email texts and routing information. KDLA would like to get at least the email part into the hMail system. For the current governor we were told that if it is permanent, the office printed it out.

Pennsylvania:

- Targeting two entities for project Governor's Office of Correspondence and PHMC's Executive Office, Barbara Franco is the Executive Director of PHMC. The Office for Information Technology is concerned about security, what kind of connections and how secure, what software. What is the "active window"? Do not want active email accessible in the Archives.
- Can pitch the intermediate records storage as being the same as a records center for paper records.
- But if the Governor's office decides to not participate, PA can test using the State Archives. PA's State Archivist has agreed to participate. The Archives currently does not receive any e-files for email.
- In some cases, messages with archival value are printed to paper. Previous governor's records are closed for 20 years. Requests for records go to the former governor's attorney.

North Carolina:

- Secretary of State's Office, Corporations Division (use Groupwise);
- Governor's Office, the Community and Citizen's Affairs Office.
- The Head of Emergency Management and the Commissioner of Insurance have also expressed an interest to records analysts that have talked about the project.
- **Chris** also suggested the recently created North Carolina Education Lottery because they have been cooperative thus far regarding writing their retention and disposition schedules.

Statistical Comparison to report back to NHPRC

Kelly: To report to NHPRC, can we get some statistics as to this is what we get in paper and this is what we get electronically? We would like to be able to say that we received x amount of

correspondence prior to electronic vs. after accepting electronic; what the time issues were for paper vs. electronic; what time was lost through records analysis.

Kentucky: We could use the paper count of printed out emails from the previous governor or commissioner or secretary and then what received in electronic format for this grant. Currently, constituent mail doesn't have significant tracking data. The office is primarily interested in when the message was received and when and how it was handled. The current governor can succeed himself, but KDLA still could get some sort of transfer for the previous term. KDLA can pitch the program to their technology folks and see what could be set up on a state server to see about getting any other agencies. Mark and his staff will be doing a lot of training in the next few months so he can bounce off the idea of participating.

Kelly: Will check with Lucy Barber at NHPRC to see what a good number of participants are. Does know that NHPRC would rather see an office high importance rather than one's own department.

Other possible opportunities:

Kelly: Met Andy Pitman of Microsoft. He's interested in possibly partnering with us and would possibly give us copies of Office 2007. **Glen** says he is very antithetical to what he says is NARA's desire to keep everything open, so he wonders just how supportive Andy will be. Adam Jansen has a good relationship with them and used to be an archivist for Microsoft. Adam has signed on with the Microsoft version of XML (which is actually more a wrapper). We can test it as a digital rights issue at the very least, which will probably be more a training aspect than actual digital rights management.

Linda: If you author an email in Word (set Word as your default email editor) then you can save it as an XML (MS version) document. But the user may not be able to do so because ITS may not allow it due to the hidden scripts within Word. Exchange 2007 can be set up to automatically go into SharePoint, which is trying to obtain DoD certification.

Glen: You can set up retention within email, but it is not obvious. It can cost a lot of money for customization to truly get a workflow and records management setup. You can set up folders to expire, but not individual items.

Glen: It might be worthwhile to see how SharePoint in the agencies mixes up different types of documents within the SharePoint system.

Q: If Exchange 2007 is no longer being saved as .pst then what are they saving it as?

A: Eml and msg. Kentucky has some agencies looking at it. Pennsylvania just starting to look at it (Archives slated to get it by end of 2008). PA standard for EDMS is Filenet but only a few agencies have it. Kentucky's email store will not be Filenet. Too expensive and hard to work with. The one KY agency that is interested in 2007 wants to put some email in it. This may be a good agency to "test" in some way with hMail.

Roles and responsibilities:

NC

- Advertise for Programmer in April for someone knowledgeable about C# on the .net platform. Glen would like to see it moved to Java by end of the project, but might not happen until next grant.

David:

- What it will take to build an installable module on a single server to work with our hMail server automatically.
- Will it work with Windows/Outlook 2000?
- What happens to the digital signature with 2003 and 2007.
- Will pull together the specs for our server and Pennsylvania's servers and go ahead and order them. Does he need to include tape backup? (Thinks it is okay).
- Ideally *David* would love to capture what the user determined to save AND everything that came in that could be compared to be able to go through the thought process of the user, including sent mail, etc.

Kelly:

- Will contact Lucy Barber about what is the ideal number of partners within each state.
- Will also explore further about testing Office 2007 and let us know.

Chris:

- Develop website about the project,, which will include grant documentation, email policies, and David's documentation about hMail.
- Chris will look at the Michigan website on their records management system implementation. They talk about the good, the bad, and the ugly of their project. What went wrong. It was suggested that Chris mirror the structure of the Michigan's site.

Kentucky:

- Can take someone's .pst file (like the head of KDLA) from one person and copy it to the hMail server to test doing that.

Pennsylvania:

- Server they are going to use isn't ready and won't be for a while. However, they are getting a server from us.
- Will send David the shipping address for the server and whether it can be a Dell or must be an IBM unit. Should it be Enterprise edition? David doesn't think so.
- Will let David know about tape backup but doesn't think it is needed.
- Will put together the training once the hardware and software is in place.

Everyone:

- If you want to start collecting mail now you can do so on a client-by-client basis. Otherwise will be the end of summer, after the new programmer has come on board.
- Everyone needs to document the time they spend on this project. Doesn't have to be per week, but it should average out. Don't need time sheets. We have to do quarterly reports.
- Programmer will work from July 2007 until June 2008.
- We need to be consistent in what we name the archival store folder for the training modules. David has used the term "Archived Mail" so far, but suggested "State Archival Mail" or SAM without the tense. But it is called Archive in Outlook's archived folder.
- Everyone send Pennsylvania what training materials they currently have for email.

- The next meeting will be at the end of September or the beginning of October.

Addendum B—Consultation Summary with Secretary of State Office (NC)

Consultation Summary—Secretary of State Office, 8/27/07, 10:00 am – 10:40 am

With Haley Haynes, Deputy Secretary of State and Bruce Garner, Chief Information Officer, SOS

Ed Southern, Druscie Simpson, and Kelly Eubank from Archives and Records Section.

Ed, Druscie, and Kelly attended a very preliminary meeting with Haley Haynes, Deputy Secretary, Secretary of State, and Bruce Garner, CIO, Secretary of State, to discuss the possibility of partnering with their office to test the email collection and preservation tool being developed by the NC Archives and Records Section. Several months ago Kelly gave Mr. Garner a copy of the technical specs from March 2007. For the meeting today, Kelly included the brief four page document drafted by David Minor as well as a rough draft of the PowerPoint presentation that will be presented at the Society of American Archivists annual meeting by Mark Myers, on August 31, 2007. These two documents are a high level overview that concisely details the tool and how we intend for it to work.

Ms. Haynes and Mr. Garner reacted very favorably to the idea of participating in the use of the tool in view of their agency's long-term storage responsibilities. Ms. Haynes mentioned the possibility of working with the Notary office as it is becoming a more prominent and important office in the SOS office. She also mentioned the Corporations office. She said that they are currently reviewing their retention schedules and this is a good time to collaborate. They will also appoint a new Chief Records Officer (currently, it is Mary Kelly). Ms. Haynes said that from her office, the most likely point of contact would be Tina Wagstaff. Bruce mentioned that the point of contact from his staff would be David Gilmore.

Ms. Haynes asked if there would be training associated with the use of email, as many people in the SOS office were unaware of their obligations regarding keeping email. Ed mentioned the workshops that are offered throughout the year as well as the possibility of presenting the workshop on-site in the SOS office. Kelly also stated that there is a training component in the grant that will be done by the Pennsylvania State that will eventually be available on-line. Kelly will notify Haley and Bruce when this has been completed.

Ms. Haynes also asked how people would know what email to archive—would that be considered part of the email grant? Kelly stated that it would work in the context of the existing records retention schedule. Those records deemed archival should be moved over, while the other email is either deleted or kept in a separate location as determined by their retention schedule. Ed acknowledged that there is an element of user appraisal concerning what email is in fact archival and of substantively archival or historical value. In practice we may end up with records that we should not have simply due to lack of time to fully evaluate every email and its importance. Thus, we may catch unimportant as well as important mail.

Mr. Garner and Ms. Haynes asked questions with regards to how the tool works. Druscie informed them that the technology works from the common Internet Message Access Protocol, (IMAP), configuration, which is a standard for transmission of email across the Internet. *For example, email stored on an IMAP server can be manipulated from a desktop computer at home, a workstation at the office, and a notebook computer while traveling, **without** the need to transfer messages or files back and forth between these computers* (source: <http://www.imap.org/>, accessed 8/27/2007 1:48pm). Using this protocol, the tool is not limited to a client-specific software tool. Druscie explained that the email preservation tool would collect both emails and attachments. It will convert the email message to XML and put an XML wrapper around the attachments. When feasible, the

attachments will be converted to PDF, but a copy of both the original message and the original attachment will be kept.

Druscie asked if SOS employees make folders in the existing Groupwise client and file their email according to folder, or if they create “buckets” for the email and drop them in accordingly. Both Haley and Bruce said it depends on the employee. For themselves, they group their email by year. Haley also commented that sometimes filing email it was could be really confusing. An email saved for one purpose might be moved to another folder because the purpose of the email had changed. Druscie noted that once email is collected with the email collection and preservation tool, it cannot be removed from its “archival” status. However, if the user moves an email to another folder, the tool will recognize that the email has been moved and move it accordingly. However, it will not accept duplicates.

Druscie stated that we would work with the SOS office regarding collection times and frequency. Bruce mentioned that the office backs up the files on the weekends, so we would not want to be collecting at the same time. Ms. Haynes will be in contact with us regarding the change in the Chief records officer and changes to their retention schedule. Kelly said she will send the web address to the project page once it went live. Both Ms. Haynes and Mr. Garner were very positive about participating.

Best wishes,
Kelly Eubank