DRAFT ACERA Meeting Minutes, 04/05/07

ADVISORY COMMITTEE ON THE ELECTRONIC RECORDS ARCHIVES
MEETING NO. 4
NATIONAL ARCHIVES BUILDING

MINUTES
DAY 2 OF 2
APRIL 5, 2007

In accordance with the provisions of Public Law 92-463, the meeting was open to the public from 9:00 a.m. to 5:00 p.m.

COMMITTEE MEMBERS

Name	Organization
Lewis Bellardo	National Archives and Records Administration
Laura E. Campbell	Library of Congress
David Carmicheal – **Not Present**	Georgia Archives
Sharon Dawes – **Not Present**	Center for Technology in Government
Luciana Duranti – **Not Present**	University of British Columbia
Dr. Richard Fennell	Administrative Office of the U.S. Courts
Daniel Greenstein – **Not Present**	University of California
Dr. Christopher Greer – Substitute for D. Atkins	National Science Foundation
Jerry Handfield	Washington State Archives
Robert Horton	Minnesota Historical Society
Dr. Robert E. Kahn	Corp. for National Research Initiatives
Andy Maltz	Academy of Motion Picture Arts and Sciences
Richard Pearce-Moses	Digital Government Information
John T. Phillips	Information Technology Decisions
Dr. Dan Reed	University of North Carolina at Chapel Hill
Adrienne Reagins	National Archives and Records Administration
Jonathan M. Redgrave	Redgrave Daley Ragan & Wagner LLP
David Rencher	Federation of Genealogical Societies
James J. Hundley, Jr. – Substitute for R. Testa	U.S. Air Force
Dr. Ken Thibodeau	National Archives and Records Administration
Allen Weinstein	National Archives and Records Administration
Dr. Kelly Woestman	Pittsburgh State University

1. Advanced Decision Support for FOIA Processing of Presidential E-Records – Dr. William Underwood

DISCUSSION OF THE PRESENTATION

What do you do about encrypted files?
Dr. Underwood – Scores have been found and need to be cracked. Currently, a password recovery tool is used.

How do you handle changes in geographic locations?
Dr. Underwood – The problem has not been solved yet but we are using historical resources to map things.

How do you categorize emails that are of a more informal nature and do you keep categorizing forever?
Dr. Underwood – We have not done email yet, experiments have mostly been on attachments. We have started to run some experiments for web pages. As for email, we don't know.

Do you create the software yourselves?
Dr. Underwood – Yes.

Is it stand-alone, or does it work in tandem with COTS software, such as word processors? Does it create alternative representations?
Dr. Underwood - Multiple copies – example of the multivalent technology. The software creates an HTML markup of a document. Access review is in prototype development stage.

Discussion continued about the challenges of ambiguity and interpretation and differentiation of the sensitive and classified information that may show up unexpected on ERA which could lead to having to bring down the entire system. Another challenge is differentiating security markings from textual references.

Lew Bellardo made the point that he thought the error rate by humans would probably be the same as the computer error rate. He also pointed out that the Intelligence community would probably be willing to live with a 10% error rate is they can see that processes and methods are protected. There was also discussion on cross contamination and how the research in this area can benefit ERA dirty word searches.

When does it [PERPOS] get integrated into ERA?
Dr Thibodeau – LMC has been briefed for technology transfer. The architecture provides a framework into which technology can be plugged in. This could happen as early as Increment 2.

The ability to process large number of objects would hinge on technology.
Dr. Underwood – We moved some of the functionality into the persistent archives prototype. We hope to demonstrate the technology in the distributed environment as a service: format conversion, file type identification, content extraction, etc.

The number of objects processed depends in large measure on the hardware available. Programs are written in Java, C++, and LISP. Not everything would run on a supercomputer. Java code would run; C++ code would require conversion; LISP would not run at all.

Andy Maltz – Are you designing it to include arbitrary file types. Images are important to my industry, not necessarily text.
Dr. Underwood – Images are not a current research topic. There are other technologies out there for images.

Is there something about your technology that would preclude a plug-in?
Dr. Underwood responds that his research deals primarily with text.

David Rencher – What is the scalability of this system? Are we confined to one space with this? It sounds very labor intensive.
Dr. Underwood – Metadata extraction is completely automated. The records are in a file system and not part of the Records Management system.

Dr. Kahn – Based on what you've seen, how long would it be to process 1 M records?
Dr. Underwood – We do not have that data yet.

DR. Kahn – Can you bound it?
Dr. Underwood – No.
Dr. Underwood further elaborated that in Bush public records testing of 5,000 records the performance was one (1) second per document, but the processing could be sped up depending on the hardware. Dr. Underwood anticipated that when further experiments are run, the performance measures would include time.

Dr. Kahn – With respect to DOD SCORM schema - Have you thought about how that might work in the context of multiple archives and presidential libraries? Can you extract metadata from one system and move it some place else? A registry of registries of meta data to share?

Dr. Underwood – This is more the kind of work that Reagan Moore is doing. My research is centered more on what NARA is focusing on. In terms of the metadata impact, we have not looked at doing that. With regard to export – most of the metadata is in containers rather than in DBMS. We do not maintain it centrally. It is in the manifest and there is a schema for that manifest structure. If there is any other metadata related to the record, it can be put in the manifest.

Andy Maltz – The manifest is not part of the record?
Dr. Underwood – No. The record is just a bit stream. Metadata can be moved into the manifest at the time of accession.

Richard Pearce-Moses – Have you looked at other types of records or content extraction tools?
Dr. Underwood – The obvious ones are Data Bases, since they are structured. "Advanced Revelation" is still around. A lot of presidential records are still in Advanced Revelation and there are no viewers for it.

Dr Thibodeau – The members of this committee would like to be actively involved in our research. Can they use PERPOS on their records?

Dr. Underwood – I will need to talk to Bob Chadduck. Our software is not open source. Data rights need to be discussed.

Dr Thibodeau – We will be happy to pursue this.

AI – To Ken Thibodeau: Find out the possibilities of the committee participating with research using PERPOS.

John Phillips – Can you clarify the relationship of your tools to e-mails. Can you do content analysis of e-mails?
Dr. Underwood – Yes.

John Phillips – Can you do the analysis of the e-mails *well*?
Dr. Underwood – No. We are running experiments to understand exactly how well we are doing. We are not attempting to understand the entire content of the natural language.

John Phillips – One would think that if you can analyze databases such as Revelation, you could do e-mails.

Richard Pearce-Moses – Given that your presentation example of the Shirley Temple Black letter was very e-mail like, why would you not be able to process e-mail?
My sense is that your software can figure out the content and form of the e-mail attachment and perhaps even the body of the e-mail.

Dr. Underwood – We have not tried to do that.

John Phillips – e-mail will require natural language processing.

Dr Thibodeau – Ten years ago we engaged John Hall (?) researchers of the presidency. But if the presidential documents are digitally pre-processed, the most significant items could be identified.

Dr. Underwood – It would be worth an experiment to scan some of the Bush records and process them through PERPOS to speed up processing.

2. General Discussion
Would it be possible to build a schedule of what happened during the presidential term to build a structured profile of a presidency? Something akin to a daily diary that could map the correspondence to a sequence of events using the context?

Dr. Weinstein – There may be some most important events that happened during a presidency that are not on the "schedule."

Richard Pearce-Moses – The professional context has shifted. An archivist now not only makes the information available, but also pre-processes it.

Dr. Weinstein – You should take into account that no one was intensely interested in the presidential records before the Richard Nixon's presidency.

There was an extended discussion regarding system performance, benchmarking, and Ken Thibodeau's handout, Performance Goals Specification, from the ERA Requirements Document (RD) that contained definitions of performance metrics. Dr. Thibodeau described how the metrics were determined and how LMC will be encouraged to meet them using award fees. Actual expectations cannot be discussed because they are being revised and are due any day.

Dr. Kahn wanted to see what the baseline is. He suggested that the committee could do a compare and contrast once the new list is available.

AI – Ken Thibodeau will get the old one to show the group and put the new one up as soon as it is available.

Jerry Handfield asked if these were NARA metrics or specifically ERA.
Ken Thibodeau answered that it is a NARA metric of ERA – things that are already online will be counted.

There was additional extended discussion of how certain metrics for records were chosen. This discussion also included what NARA's expectations of LMC are and how NARA will know if LMC is meeting them. There was also discussion of testing and how it fits in with managing NARA's expectations of LMC.

Lunch break

3. Subcommittee Topic Discussion
Following the lunch break there were discussions on possible topics for subcommittee discussions.

One possibility is how NARA is going to frame requirements for other agencies to transmit data. Agencies are keeping and transferring data in a variety of formats and NARA is required to accept records. Who is in control of this process?

Dr Thibodeau suggested that one way to gain more understanding of ERA requirements would be to review the Vivid Description and Performance Goal Specifications, as well as NARA Agency Goals.

Dr. Kahn – What was the baseline?

Dr Thibodeau – The biggest challenge in measuring the improvement is the absence of a baseline. Most processes were not measured.

Richard Pearce-Moses – I can echo that for our archives.

Jerry Handfield – Is the percent improvement measure applied against what NARA or ERA must achieve?

Dr Thibodeau – NARA. And there we do have a baseline.

Dr. Kahn agreed that most of these metrics are NARA metrics and reiterated his desire to understand what burden with respect to system performance was placed on Lockheed Martin? What part of it did they have to do?

Jerry Handfield said that in his experience the first major problem was the rate at the point of ingest.

Dr. Kahn – What was Lockheed Martin asked to do? You need to bring specific requirements, especially for the Presidential Libraries. We would like to see the Performance Document and Performance Model.

Dr Thibodeau – The documents are still under development.

John Phillips – You need concrete requirements, or the contractor will point back to the requirements document and scream "Scope Change!"

DISCUSSION proceeded to choosing the topics for the subsequent sessions and policy for standards. How is NARA going to deal with the future requirements at competing levels? Who is calling the shots?

- Standards vs. Improvisation as necessary. Is there only one way?
- NARA does not want to be in the role of standards developer.
- Dr. Weinstein's conclusion is that NARA may need to improvise until such time that it could formulate the desired standards and seek the legislation to enforce them.

Would standards be a topic for future discussion?

Dr Thibodeau – Having record format requirements in the past constrained what NARA could process. Between 1972 and 1990 NARA processed primarily flat files, such as socio-economic data.

Jerry Handfield – NARA is not likely to get compliance unless the standard is inherently popular.

Dr Thibodeau – Should we explore a policy for standards?

Dr. Kahn – Government should not normally be in the business of standards development, other than to meet internal needs of the Government itself

John Phillips – What relationship does ERA have with NIST?

Dr Thibodeau – We have an active relationship with NIST. Standards coming out of the geographic mapping community, such as ESRI, actually work. Content standards are of greatest value because they can reduce the variability of the metadata to deal with. The standard produced by the Federal Geographic Data Committee (FGDC), The Content Standard for Digital Geospatial Metadata (CSDGM), that is the US Federal Metadata standards and is often referred to as the FGDC Metadata Standard is rapidly becoming a national standard.

Andy Maltz – Electronic archiving is enterprise-wide issue that fundamentally changes everything. In order to execute certain things, though, they must be mandated from above.

Dr. Kahn – Another issue is the timeliness of submissions to the archives.

Jerry Handfield – Our policy is to transfer permanent records into the archives upon creation, even though they may not be available immediately.

Dr. Kahn – Any organization could have its own archive of permanent records. They can be packaged and zipped and shipped to NARA at a scheduled date.

Dan Reed – There is the "Tom Sawyer" approach – convincing the archives that they have something to gain from transferring their records to the archives in a timely manner. Budgets may be an incentive.

Dr Thibodeau – The records coming to NARA are available immediately. But archivists do not have enforcement power.

Dr. Kahn - Auditors do not have to sell anyone their services.

Richard Pearce-Moses – [I am] cynical in terms of compliance. But I think that technology evolution may drive the adoption of uniform practices – XML will do a lot towards that.

Dr. Kahn – Shall we pursue a discussion of the policy for standards?
VOTE:
SUMMARY – Six (6) votes for and two (2) votes against.

Dr. Kahn – Assuming material is available in ERA, what about other systems' ability to interact with ERA?

Dr Thibodeau – Yes, it would be very valuable for the external systems to be able to interact with ERA.

Dr. Kahn – Information about the Handle System is available at http://www.handle.net/

OTHER TOPICS

Dr Thibodeau – What could the rest of the world do when ERA is available?

Richard Pearce-Moses – map shops [would benefit]

Andy Maltz - 20-somethings could be put on the job.

John Phillips – Would you want the contents of the archive popping up on Google? Look and feel of the public interface would make or break public interaction.

Jerry Handfield – Would there be amazon.com – like features?

Dr Thibodeau – There would have to be, otherwise you wouldn't be getting 'records', you would be getting 'documents.'

John Phillips – What happened to the consortium on Digital Libraries Initiative?

Dr Thibodeau – I am NARA's representative to the Digital Library Initiative.

Dr. Kahn – CrossRef in Massachusetts is a service of the Publishers International Linking Association (PILA). Something akin to CrossRef's deal with Google can probably be arranged for ERA.

Reference – CrossRef [http://www.crossref.org] is the citation-linking backbone for online publications. Established in 2000 by scholarly publishers as an independent, non-profit entity, it enables researchers to navigate electronic journals, across publishers, based on open-standards technology (the Digital Object Identifier, or DOI, system).

Jerry Handfield – Should Dr. Weinstein demand that the records be indexed before being deposited at NARA?

DISCUSSION on indexing and access.

The committee members advocate indexing after accession.

Richard Pearce-Moses expresses excitement about Dr. Underwood planned visit to Arizona.

4. ADJOURNMENT

The meeting adjourned at 5:00 p.m.

I herby certify that, to the best of my knowledge, the foregoing minutes are accurate and complete.

Adrienne M. Reagins
Secretariat
Advisory Committee on the Electronic Records Archives

Robert Kahn, Ph.D.
Chairman
Advisory Committee on the Electronic Records Archives

These minutes will be formally considered by the Committee at its next meeting, and any corrections or notations will be incorporated in the minutes of that meeting.