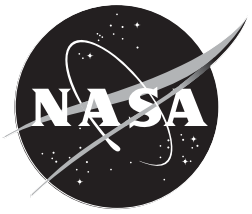
NASA/CP—2004–212750



NASA/IEEE MSST 2004
Twelfth NASA Goddard Conference on Mass Storage Systems and Technologies

in cooperation with the

Twenty-First IEEE Conference on Mass Storage Systems and Technologies

Edited by
Ben Kobler, Goddard Space Flight Center, Greenbelt, Maryland
P C Hariharan, Systems Engineering and Security, Inc., Greenbelt, Maryland

Proceedings of a conference held at
The Inn and Conference Center
University of Maryland, University College,
Adelphi, Maryland, USA
April 13-16, 2004

National Aeronautics and
Space Administration

Goddard Space Flight Center
Greenbelt, Maryland 20771

April 2004

The NASA STI Program Office … in Profile

Since its founding, NASA has been dedicated to the advancement of aeronautics and space science. The NASA Scientific and Technical Information (STI) Program Office plays a key part in helping NASA maintain this important role.

The NASA STI Program Office is operated by Langley Research Center, the lead center for NASA's scientific and technical information. The NASA STI Program Office provides access to the NASA STI Database, the largest collection of aeronautical and space science STI in the world. The Program Office is also NASA's institutional mechanism for disseminating the results of its research and development activities. These results are published by NASA in the NASA STI Report Series, which includes the following report types:

- TECHNICAL PUBLICATION. Reports of completed research or a major significant phase of research that present the results of NASA programs and include extensive data or theoretical analysis. Includes compilations of significant scientific and technical data and information deemed to be of continuing reference value. NASA's counterpart of peer-reviewed formal professional papers but has less stringent limitations on manuscript length and extent of graphic presentations.
- TECHNICAL MEMORANDUM. Scientific and technical findings that are preliminary or of specialized interest, e.g., quick release reports, working papers, and bibliographies that contain minimal annotation. Does not contain extensive analysis.
- CONTRACTOR REPORT. Scientific and technical findings by NASA-sponsored contractors and grantees.
- CONFERENCE PUBLICATION. Collected papers from scientific and technical conferences, symposia, seminars, or other meetings sponsored or cosponsored by NASA.
- SPECIAL PUBLICATION. Scientific, technical, or historical information from NASA programs, projects, and mission, often concerned with subjects having substantial public interest.
- TECHNICAL TRANSLATION. English-language translations of foreign scientific and technical material pertinent to NASA's mission.

Specialized services that complement the STI Program Office's diverse offerings include creating custom thesauri, building customized databases, organizing and publishing research results . . . even providing videos.

For more information about the NASA STI Program Office, see the following:

- Access the NASA STI Program Home Page at http://www.sti.nasa.gov/STI-homepage.html
- E-mail your question via the Internet to help@sti.nasa.gov
- Fax your question to the NASA Access Help Desk at (301) 621-0134
- Telephone the NASA Access Help Desk at (301) 621-0390
- Write to:
  NASA Access Help Desk
  NASA Center for AeroSpace Information
  7121 Standard Drive
  Hanover, MD 21076-1320

Available from:

NASA Center for AeroSpace Information
7121 Standard Drive
Hanover, MD 21076-1320
Price Code: A17

National Technical Information Service
5285 Port Royal Road
Springfield, VA 22161
Price Code: A10

NARA's ELECTRONIC RECORDS ARCHIVES (ERA) – THE ELECTRONIC RECORDS CHALLENGE

Mark Huber
American Systems Corp.
National Archives and Records Administration
8601 Adelphi Rd., Rm. 1540, College Park, MD 20740
Tel: +1-301-837-0420
mark.huber@nara.gov

Alla Lake
Lake Information Systems, LLC
National Archives and Records Administration
8601 Adelphi Rd., Rm. B550, College Park, MD 20740
Tel: +1-301-837-0399
alla.lake@nara.gov

Robert Chadduck
National Archives and Records Administration
8601 Adelphi Rd., Rm. 1540, College Park, MD 20740
Tel: +1-301-837-0394
Robert.chadduck@nara.gov

Abstract
The National Archives and Records Administration (NARA) is the nation's recordkeeper. NARA is a public trust that safeguards the records of the American people, ensuring the accountability and credibility of their national institutions, while documenting their national experience. Today NARA holds an estimated 4 billion records nationwide. The Archives consists of the permanently valuable records generated in all three branches of the Federal Government. These record collections span this country's entire experience, across our history, the breadth of our nation, and our people. While paper documents presently predominate, NARA holds enormous numbers of other media, such as reels of motion picture film, maps, charts, and architectural drawings, sound and video recordings, aerial photographs, still pictures and posters, and computer data sets. It is that last medium, the electronic records, that is the fastest growing record keeping medium in the United States and elsewhere in the world. Since 1998, NARA has established key partnerships with Federal Agencies, state and local governments, universities, other national archives, the scientific community, and private industry to perform research enabling better understanding of the problems and the possibilities associated with the electronic records challenge. The challenge of electronic records encompasses the proof and assurance of records authenticity and assurance of record persistence and ready access to records over time.

1. Background/General Project Description

"Electronic records pose the biggest challenge ever to record keeping in the Federal Government and elsewhere. There is no option to finding answers…the alternative is irretrievable information, unverifiable documentation, diminished government accountability, and lost history."

John Carlin, The Archivist of the United States

The National Archives and Records Administration (NARA) is the nation's recordkeeper. NARA is a public trust that safeguards the records of the American people, ensuring the accountability and credibility of their national institutions, while documenting their national experience. Pursuant to legislation codified under Title 44 of the United States Code the Archivist of the United States has authority to provide guidance direction and assistance to Federal officials on the management of records, to determine the retention and disposition of records, to store records in centers from which agencies can retrieve them, and to take into the archival facilities of the National Archives and Presidential libraries, for public use, records that he determines "to have sufficient historical or other value to warrant their continued preservation by the United States Government." (44 U.S.C. 2107) Similarly, under the Presidential Records Act, when a President leaves office, the Archivist of the United States assumes responsibility "for the custody, control, and preservation of, and access to, the Presidential records of that President". Both the Government and the public rely on NARA to provide this and subsequent generations of the American public with access to extraordinarily high accretion rate, increasingly diverse, and arbitrarily complex collections of historically valuable federal, presidential and congressional electronic records collections.

The technology challenge confronting NARA is repeatedly confirmed as among the President's research priorities. In the supplement to the President's budget for fiscal year 2004, The National Science and Technology Council expressly acknowledges that "R&D in advanced technologies that enable preservation and utility of electronic information archives…," and "…digital archives of core knowledge for research and learning" is "far from finished." Especially prominent is the Council's explicit identification of "….substantial technical issues – such as interoperability among file formats, indexing protocols, and interfaces; data management, storage and validation; … and long term preservation – that impede development of digital libraries…" Similarly noted is research enabling agencies to move "…toward two ambitious goals: quick, easy, and secure on-line access for citizens to government services and information, and radical reduction in internally duplicative record-keeping, ... through coordinated development of IT standards and procedures..." [1]

Experts predicted in FY2003 that electronic records volumes will swell by orders of magnitude over this decade, presenting enormous challenges for society along with unprecedented opportunities for U.S. advanced research and technological innovation.", …fused with requirements for… "technologies for rapid mining, filtering, correlating and assessing of vast quantities of heterogeneous and unstructured data", and… "tools for collecting, archiving and synthesis." [2]

Similarly, among the president's FY2002 research priorities:

"Strategies to assure long-term preservation of digital records constitute another particularly pressing issue for research. As storage technologies evolve with increasing speed to cope with the growing demand for storage space, the obsolescence of older storage hardware and software threatens to cut us off from the electronically stored past." [3]

The Archivist is authorized by law to "conduct research with respect to the improvement of records management practices and programs." [44 U.S.C Section 2904(c)(2)]. Since 1998, NARA has established key partnerships with Federal Agencies, state and local governments, universities, other national archives, the scientific community, and private industry to perform research enabling better understanding of the problems and the possibilities associated with the electronic records challenge.

NARA's Key Research Partners

- National Science Foundation (NSF)
- San Diego Supercomputer Center (SDSC)
- University of Maryland (UMd)
- Georgia Tech Research Institute (GTRI)
- U.S. Army Research Laboratory (ARL)
- National Computational Science Alliance (NCSA)
- National Institute of Standards of Technology (NIST)
- National Nuclear Security Administration (NNSA)
- National Aeronautics and Space Administration (NASA)
- U.S. Department of Defense (DOD)
- Library of Congress (LC)
- International Research on Permanent Authentic Records in Electronic Systems (InterPARES)
- Digital Library Federation (DLF)
- Global Grid Forum (GGF)

NARA's ERA Program includes ongoing sponsorship, support, and collaboration in technology research activities relevant to developing and sustaining the systematic capability for transfer, preservation, and sustained access to electronic records. ERA must be dynamic in response to continuing technology evolution, ensuring that electronic records delivered to future generations of Americans are as authentic decades in the future as they were when first created.

Among the findings presented in the report of the Committee on Digital Archiving and the National Archives and Records Administration of the Computer Science and Telecommunications Board (CSTB) for the National Research Council of the National Academies are findings that while no turnkey system, application, or product exists in the marketplace which meets NARA's requirements, the system can and should be built. [4]

2. Program Status

In response to the digital records challenge, Congress, in November 2001, acting through the Treasury and General Government Appropriations Act {P.L.107-67}, approved the fiscal 2002 budget that included $22.3 million for Electronic Records Archives (ERA) Program. Similarly, in January 2003, Congress, acting through the Consolidated Appropriation Resolution, 2003 {P.L.108-7}, approved the fiscal 2003 budget that included $11.8 million for the Electronic Records Archives (ERA) Program. At the time of this writing, and while the final appropriations have not passed, both the House of Representatives and the Senate have agreed to fund the ERA Program at the $35.7M level in the President's FY2004 request. The official Request for Proposal (RFP) for the ERA system was released to the public on December 5, 2003. At the time of the RFP release, proposals from industry were required to be submitted to NARA by January 28, 2004. The ERA program schedule calls for up to two contract awards to be made by mid-2004.

NARA has structured the ERA procurement to fundamentally be a challenge to industry to propose innovative ways to address the challenges represented by the large number and variety of electronic records generated and used by the Federal government. The ERA procurement strives to define the electronic records challenge without prescribing implementations or techniques with which to address the issues. Again, NARA wants to engage industry in crafting long term responses to the various technical and operational issues that ERA represents. This paper goes on to explore some of the archival, technical, and operational issues that the ERA program sees as important to the success of ERA.

3. Goals, Issues, and Challenges for Electronic Records - Persistence, obsolescence, access over time

Today NARA holds in the National Archives of the United States and the Presidential Libraries an estimated 4 billion records nationwide. The archives consist of the permanently valuable records generated in all three branches of the Federal Government, supplemented with donated documentary materials. [5]

These records span this country's entire experience, across our history, the breadth of our nation, and our people. Not surprisingly, with the passage of time, the medium of the records of the United States has become diverse in format. While paper documents presently predominate, NARA holds enormous numbers of

- reels of motion picture film,
- maps, charts, and architectural drawings,
- sound and video recordings,
- aerial photographs,
- still pictures and posters, and
- computer data sets. [6]

It is that last medium – computer data sets - the electronic records, that is the fastest growing record keeping medium in the United States and elsewhere in the world. According to the *How Much Information? 2003* study from the University of California

Berkley School of Information Management and Systems, released in October 2003, the worldwide production of original information stored digitally on magnetic media has grown by 80% in the time elapsed between the 1999 and 2002 samples. The upper boundary study volume estimate in that category of information for 1999 was 2.8 Peta Bytes and for 2002 – 4.99 Peta Bytes. [7]

The digital (electronic) storage of information –has been growing in proportion to the rise in creation and use of information in general. There is no consensus optimal method for the long term preservation of electronic records. A number of approaches are being used in the industry singly and in combination. Each of the approaches brings with it its own cost, as well as operational and reliability concerns. The larger the size of the electronic records holdings, the more important it is to carefully select and design the preservation approach.

Preserving electronic records serves the same fundamental *purpose* as preserving any other type of record: to enable the records to continue to provide evidence and information about the decisions, acts, and facts described in the records with the same degree of reliability as when the record was created. However, the *process* of preserving electronic records is substantially different than the preservation of traditional, non-electronic records. Traditional records are aptly termed "hard copy" in that the information that the record contains is inscribed in a hard, indissoluble manner on a physical medium, and the physical inscription conveys the information the record is intended to provide. Therefore, preservation traditionally focused on the physical object. However, an electronic record is inscribed on a physical medium as a sequence of binary values which must always be translated into a different form – the form of a record – in order to communicate the information the record was meant to convey. Therefore, preserving an electronic record requires maintaining the ability to reproduce that record from stored data. While the preservation of a paper record can be deemed successful if that record remains physically intact in storage, the success of a process of preserving an electronic record can only be verified by translating the stored bits into the form of the record. It is the result of this reproduction, not the stored bits, that literally **is** the electronic record. If the wrong process is applied, or if the process is not executed correctly, the result will not be an authentic copy of the record. Over time, reproducing an electronic record is challenging because the conventions for representing information in digital form change along with hardware and software. Newer systems may not be able to process older formats, or may do so incorrectly. [8]

Archiving of electronic records brings with it an increased challenge of authenticity of the record and a more difficult burden of proof of that authenticity. For the ERA program, electronic record *authenticity* is defined as *the property of a record that it is what it purports to be and has not been corrupted.* Given the legal, historical, and cultural significance of national or institutional record holdings, authenticity of the records is essential. Establishing authenticity of a paper, photographic negative, or other physical medium-based record has historically been accomplished by establishing that the record itself is, or is based on, an original via the proof that the medium of the record (or the medium of the basis record) is the original and there is a clear chain of custody

associated with the record. Stringent requirements assigned to electronic record collections to support a continuing burden of proof relates to the attainment of criteria for authenticity over time. Electronic records present special challenges with respect to the proof of record authenticity as the record is preserved over time due to the both increased risk of corruption of the record when it exists in digital form.

Records are being created in progressively larger volumes through the use of electronic hardware and the associated software applications. Some of the records are traditional textual or graphic documents that could have been originated with the use of pen and paper. At least in theory, their content can be preserved in hard copy. The bulk of them, however, are in most respects indelibly tied to the technology that produced them, such as the contents of data base systems, interactive Web pages, geographic information systems, and virtual reality models. [9] These later types of electronic records need to be preserved in electronic form in order to preserve the essential properties of the record other than pure content - the context, structure, and behavior. Whenever a mix of technologies are involved in the creation, maintenance, and presentation of the record, preservation is far more involved than the preservation of the precise sequence of bits constituting instrument reading, an ASCII text, or a bitmap graphic document.

All of the electronic records in lesser or greater degrees rely for access on the use of technologies that arise and evolve rapidly and just as rapidly become obsolete. Computing platforms on which the records are created, preserved, or examined, communication infrastructures interconnecting these platforms, data recording media, and, perhaps most importantly, data recording formats are all subject to rapid obsolescence while the records themselves must persist.

Preservation approaches for electronic records are multifold and can be broadly categorized in following areas of concern:

- Media
- Hardware Technology
- Software Technology, including record formats
- Archival: provenance, authenticity, context, structure and appearance.

A significant complicating factor in preservation of an electronic record is the necessity to preserve some of the associated linkages to other records. The loss of such linkages may, at best, lead to the loss of context or, at worst, render the record itself unreadable.

Preservation of electronic records is the end-to-end process which enables re-production of an authentic copy of the record. To assure that reproduction preservation of electronic records extends beyond protection of the record physical medium to protection of record accessibility and assuring record authenticity over time. Assuring persistence of records means ensuring that the records are not only readable but also intelligible after the passage of time. Assuring record authenticity means ensuring that the records are not inadvertently or deliberately altered or corrupted over time and that the authenticity can be adequately proven. [10]

Finally, a fundamentally important aspect of ensuring that the records are accessible over time is appropriate processing of the records as they enter the electronic archive with respect to establishing appropriate searchable archival structures and relationships and extraction and storage of associated metadata. Preservation methods which maintain dependencies of records to obsolete technologies tend to increasingly constrain access over time. Continuing general public access to old and obsolete technologies may not be possible except in highly limited environments or circumstances. In example, general public access to an emulator appropriate to enable reliable future rendition of electronic records created in the technology of an early 1990's proprietary geographic information system in the hypothetical context of 2020 vintage computing is not presently assured.

4. System Characteristics and Drivers

The ERA system, because of its size, scope, ingest and access loads, and commitment to long term preservation and servicing of government records, will require deployment and design approaches that support its unique nature and mission goals. When designing and deploying ERA, NARA must take a long term view for the system's operation and its required scalability, reliability, and cost effective operations. This long term vision will accommodate the use of outsourcing of potential processing and hosting services while at the same time ensures NARA's stewardship of the records trusted to it.

4.1 Design and Deployment Goals

There are certain assumptions and drivers that sculpt the deployment approach for ERA. These assumptions and design drivers are collectively considered the design and deployment goals for the ERA program. These goals include:

- NARA must own and control at least one set of all holdings of electronic records entrusted to it. This is required for protection of the records and fulfillment of NARA's mission to ensure long term preservation and access to the government's records.
- The ERA system is one of NARA's contributions to the Federal Enterprise Architecture (FEA) and fulfills a critical role in the development and deployment of NARA's own Enterprise Architecture (EA).
- The design and deployment vision for ERA must allow for the contracting out of record processing and access support, if NARA chooses to exercise that option in the future. The contracting out of record services must be done within the context of NARA's mission and ultimate responsibility for the integrity of the records.
- Minimize government ownership of equipment and facilities. This desire must be balanced against NARA's stewardship of the records and commitment to FEA support.
- Allow industry and academia to provide value added services on record holdings.
- Produce a highly reliable system design. Characteristics of such a design include:
    - Avoidance of single point/site of failure.
    - Graceful performance degradation of the system when failures occur
    - Maintain system operations in face of remedial maintenance (RM), preventative maintenance (PM), and planned upgrades/changes

4.2 System Design Drivers

In addition to the deployment goals, the design and deployment of ERA must take into account certain architectural demands and aspects of the ERA record preservation domain itself. These drivers must be accommodated in any deployment and design strategy for the ERA system. These drivers include:

- The size of the ERA record holdings. ERA permanent records holdings are projected to be in excess of 100 PBs of data 12-15 years after deployment, with continued growth in holdings in subsequent years. The sheer volume of data that must be accommodated, as well as its associated access loads, is a huge driver that must be accounted for in the ERA architecture. Architectural concepts including distributed deployment(s), load balancing techniques, and multiple sources for access to high demand records are applicable to the holdings size aspect of ERA.
- Insuring the integrity of the record holdings. The records must be protected from loss, alteration, or the lack of access capability over time. Appropriate security and accommodation of timely backup of holdings with subsequent restoration of access are techniques that are required in this area.
- The evolutionary nature of the ERA system. This aspect is most pronounced in two areas:
  - Changes to the Persistent Preservation approaches used for records. Over time electronic records will need to be stored, represented, and accessed in different ways given the forward march of computer technology and the rapid obsolescence of formats and techniques.
  - Independent of the record preservation techniques, the general infrastructure and support technologies used in ERA will need to be updated and upgraded over time. Technology insertion into the ERA design will be imperative.
- The heterogeneity of assets in ERA will complicate storing and providing access to the assets, as well as preserving them. The scope of this issue can be appreciated by considering that ERA records can be classified via three different attributes.
  - Record Types (RTs) – Any record will be classified according to its intellectual format. Examples of record types include letters, ledgers, maps, reports, etc.
  - Data types (DTs) – A data type is a set of lexical representations for a corresponding set of values. The values might be alphabetic characters, numbers, colors, shades of grey, sounds, et al. The lexical representation of such values in digital form assigns each value to a corresponding binary number, or string of bits. A data type may be simple, such as the ASCII representation of alphabetic characters, or composite; that is, consisting of a combination of other data types. An electronic record consists of one or more digital components; that is, strings of bits each of which has a specific data type.
  - Varied classes/collections of holdings – Records of the same RT and DT may still belong to different record series or collections, which further define the nature of the record. Examples of high level collections or

series could include particular Presidential collections, Federal record series, and potentially record series in Federal Record Centers (FRCs).

5. Conclusion

This paper has provided an overview of the NARA ERA program and the challenges that face NARA is the areas of electronic records preservation, system deployment, and archival management of the Nation's permanent records. The NARA ERA program represents a bold initiative in electronic records management and preservation and is a call for industry to propose new and innovative approaches to the unique issues NARA faces as the steward of the government's electronic records. Through the fusion of different technologies such as distributed computing, large scale object storage and access methods, secure infrastructure, and forward thinking record preservation strategies, ERA will open and exciting new era for electronics records management and access.

References:

[1] The Networking and Information Technology, R&D (NITRD), SUPPLEMENT TO THE PRESIDENT'S BUDGET FOR FY 2004, A Report by the Interagency Working Group on Information Technology Research and Development National Science and Technology Council, September 2003.

[2] The Networking and Information Technology, R&D (NITRD), SUPPLEMENT TO THE PRESIDENT'S BUDGET FOR FY 2003, A Report by the Interagency Working Group on Information Technology Research and Development National Science and Technology Council, July 2002

[3] The Networking and Information Technology, R&D (NITRD), SUPPLEMENT TO THE PRESIDENT'S BUDGET FOR FY 2002, A Report by the Interagency Working Group on Information Technology Research and Development National Science and Technology Council, July 2001

[4] Building an Electronic Records Archives and Records Administration: Recommendation for Initial Development, http://www.nap.edu/openbook/0309089476/html/R1.html

[5] http://www.archives.gov/about_us/reports/2002_annual_report_measuring_success.pdf

[6] NARA's Strategic Directions for Federal Records Management, July 31, 2003, http://www.archives.gov/records_management/pdf/strategic_directions.pdf, referenced 10/2003.

[7] *How Much Information? 2003,* http://www.sims.berkeley.edu/research/projects/how-much-info-2003/execsum.htm#summary

[8] *Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years*, Kenneth Thibodeau, July 10, 2002

[9] *Preserving the Long-Term Authenticity of Electronic Records: The InterPARES Project*, Heather MacNeil, University of British Columbia AABC Newsletter, Volume 10 No. 2 Spring 2000

http://aabc.bc.ca/aabc/newsletter/10_2/preserving_the_long.htm
[10] The Long-Term Preservation of Authentic Electronic Records: Findings of the InterPares Project. 2003. http://www.interpares.org/book/index.cfm & http://www.archives.gov/electronic_records_archives/pdf/preservation_and_access_levels.pdf