Building an Electronic Records Archive at the National Archives and Records Administration: Recommendations for a Long-Term Strategy

Robert F. Sproull and Jon Eisenberg, Editors, Committee on Digital Archiving and the National Archives and Records Administration, National Research Council

ISBN: 0-309-55181-1, 112 pages, 8 1/2 x 11, (2005)

Building an Electronic Records Archive at the National Archives and Records Administration

Recommendations for a Long-Term Strategy

Committee on Digital Archiving
and the National Archives and Records Administration

Computer Science and Telecommunications Board

NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES

Robert F. Sproull and Jon Eisenberg, Editors

THE NATIONAL ACADEMIES PRESS
Washington, D.C.
www.nap.edu

THE NATIONAL ACADEMIES PRESS 500 Fifth Street, N.W. Washington, DC 20001

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the committee responsible for the report were chosen for their special competences and with regard for appropriate balance.

Support for this project was provided by the National Archives and Records Administration under Contract No. NAMA-02-C-0012. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the organizations or agencies that provided support for the project.

International Standard Book Number 0-309-09696-0

Cover design by Jennifer M. Bishop.

Copies of this report are available from the National Academies Press, 500 Fifth Street, N.W., Lockbox 285, Washington, DC 20055; (800) 624-6242 or (202) 334-3313 (in the Washington metropolitan area); Internet, http://www.nap.edu.

Printed in the United States of America

THE NATIONAL ACADEMIES

Advisers to the Nation on Science, Engineering, and Medicine

The **National Academy of Sciences** is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Bruce M. Alberts is president of the National Academy of Sciences.

The **National Academy of Engineering** was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. Wm. A. Wulf is president of the National Academy of Engineering.

The **Institute of Medicine** was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Harvey V. Fineberg is president of the Institute of Medicine.

The **National Research Council** was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Bruce M. Alberts and Dr. Wm. A. Wulf are chair and vice chair, respectively, of the National Research Council.

www.national-academies.org

COMMITTEE ON DIGITAL ARCHIVING AND THE NATIONAL ARCHIVES AND RECORDS ADMINISTRATION

ROBERT F. SPROULL, Sun Microsystems, *Chair*
HOWARD BESSER, University of California, Los Angeles
JAMIE CALLAN, Carnegie Mellon University
CHARLES DOLLAR, Dollar Consulting
STUART HABER, Hewlett-Packard Laboratories
MARGARET HEDSTROM, University of Michigan
MARK KORNBLUH, Michigan State University
RAYMOND LORIE, IBM Almaden Research Center
CLIFFORD LYNCH, Coalition for Networked Information
JEROME H. SALTZER, Massachusetts Institute of Technology
MARGO SELTZER, Harvard University
ROBERT WILENSKY, University of California, Berkeley

Staff

JON EISENBERG, Study Director and Senior Program Officer
STEVEN WOO, Program Officer (through August 2004)
DAVID PADGHAM, Research Associate
JENNIFER M. BISHOP, Senior Project Assistant

Preface

Just as its constituent agencies and other organizations do, the federal government generates and increasingly saves a large and growing fraction of its records in electronic form. Recognizing the ever-greater importance of these electronic records for its mission of preserving "essential evidence," the National Archives and Records Administration (NARA) launched a major new initiative, the Electronic Records Archives (ERA) initiative, in 1998. NARA subsequently requested that the National Research Council's Computer Science and Telecommunications Board conduct a two-phase study to provide NARA with advice as it develops the ERA program.

Phase one of the study resulted in the preparation of two reports by the Committee on Digital Archiving and the National Archives and Records Administration. The committee's first report, *Building an Electronic Records Archive at the National Archives and Records Administration: Recommendations for Initial Development*,[1] focused on design and engineering issues related to NARA's Electronic Records Archives program. Also, the committee issued a letter report[2] in October 2003 that elaborated on issues discussed in its first report, tying these issues specifically to NARA's draft request for proposals for the ERA. Although some of the conclusions in these two reports relate to specific development initiatives and early design ideas, most of the

[1]National Research Council. 2003. *Building an Electronic Records Archive at the National Archives and Records Administration: Recommendations for Initial Development,* Robert F. Sproull and Jon Eisenberg (eds.). The National Academies Press, Washington, D.C. The "Summary and Recommendations" chapter of the 2003 study is reprinted in Appendix B of the present report.

[2]National Research Council. 2003. "Letter Report on Building an Electronic Records Archive at the National Archives and Records Administration." The National Academies Press, Washington, D.C., October 16. The letter report is reprinted in Appendix C of this report.

observations about archive system design are not tied to these specifics and are intended to remain useful to NARA as it develops, refines, and iterates the ERA program.

This final report is the result of phase two of the study. Prepared by the same committee (see Appendix D for biographical information), it examines longer-term, more strategic issues related to electronic records archiving at NARA. These issues include technology and other trends that shape the context in which the ERA exists, the archival processes of the ERA itself, and the future evolution of the ERA system. This final report also discusses record integrity, which the committee's first report did not address in detail.

The committee thanks the many people who made this report possible, although of course responsibility for the final result is its own. The support and assistance of the ERA program staff, especially Kenneth Thibodeau, Robert Chadduck, and Richard Steinbacher, are greatly appreciated. A number of individuals from NARA, other federal agencies, and the private sector, listed in Appendix A, provided valuable input to the committee during the course of its work. Jennifer M. Bishop, CSTB program associate, facilitated our work throughout the course of this project. David Padgham conducted background research and made a number of contributions to the committee's reports.

Robert F. Sproull, *Chair*
Committee on Digital Archiving and the National
Archives and Records Administration

Acknowledgment of Reviewers

This report has been reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise, in accordance with procedures approved by the National Research Council's Report Review Committee. The purpose of this independent review is to provide candid and critical comments that will assist the institution in making its published report as sound as possible and to ensure that the report meets institutional standards for objectivity, evidence, and responsiveness to the study charge. The review comments and draft manuscript remain confidential to protect the integrity of the deliberative process. We wish to thank the following individuals for their review of this report:

William Y. Arms, Cornell University,
Paul Conway, Duke University,
W. Bruce Croft, University of Massachusetts, Amherst,
Hector Garcia-Molina, Stanford University,
James Gray, Microsoft Bay Area Research Center,
Michael E. Lesk, Rutgers University,
Charles McClure, Florida State University,
Mark Seiden, MSB Associates, and
J. Timothy Sprehe, Sprehe Information Management Associates.

Although the reviewers listed above provided many constructive comments and suggestions, they were not asked to endorse the conclusions or recommendations, nor did they see the final draft of the report before its release. The review of this report was overseen by Robert

J. Spinrad, Xerox Corporation (retired). Appointed by the National Research Council, he was responsible for making certain that an independent examination of this report was carried out in accordance with institutional procedures and that all review comments were carefully considered. Responsibility for the final content of this report rests entirely with the authoring committee and the institution.

Contents

Summary and Recommendations

The National Archives and Records Administration (NARA) launched its Electronic Records Archives (ERA) program in 1998 to create a system to preserve and provide access to federal electronic records. Early steps in the ERA program included NARA's exploration of possible solutions and its undertaking of development projects with partners, including the San Diego Supercomputer Center, the University of Maryland, and the Georgia Tech Research Institute. In 2004, NARA released the final version of its ERA request for proposals (RFP) and selected two contractors to develop designs for the ERA.

The first report[1] of the National Research Council's Committee on Digital Archiving and the National Archives and Records Administration and its subsequent letter report[2] provided recommendations on design, engineering, and related issues facing the ERA program, which was being conceived at the time of their writing. Although some of the reports' conclusions relate to specific development initiatives and early design ideas, most of the observations about archive system design are not tied to these specifics and should remain useful to NARA as it develops, refines, and iterates the ERA program. The "Summary and Recommendations" chapter of the committee's first report is reprinted in Appendix B of the present report, and its 2003 letter report is reprinted in Appendix C.

The ERA program has been involved in procurement activities, but this committee has not been given access to information about the ERA design since the issuance of the ERA program's RFP, and this report does not contain—or comment on—technical details about the current system designs. The committee does urge that NARA find a mechanism, consistent with

[1]National Research Council. 2003. *Building an Electronic Records Archive at the National Archives and Records Administration: Recommendations for Initial Development,* Robert F. Sproull and Jon Eisenberg (eds.). The National Academies Press, Washington, D.C.

[2]National Research Council. 2003. "Letter Report on Building an Electronic Records Archive at the National Archives and Records Administration." The National Academies Press, Washington, D.C., October 16.

procurement regulations, for obtaining a comprehensive outside technical review of the ERA designs that are proposed by its contractors. The committee's first report recommended that NARA establish an ongoing advisory group on digital preservation and information technology (IT) system design; such a body might be charged with carrying out a review of this type.

This final report focuses on longer-term, more strategic issues that are related to electronic records archiving at NARA, including technology and other trends that shape the context in which the ERA exists, the archival processes of the ERA itself, and the future evolution of the ERA system. In addition, it addresses an important set of technical and design issues associated with assuring record integrity and authenticity that were not covered in detail in the committee's earlier reports.

NARA recognizes that it faces a significant challenge. The ERA program, the Records Management Redesign initiative, and NARA's work on records management under the e-government initiative led by the Office of Management and Budget all reflect NARA's awareness of these problems and represent positive steps in starting to address them. As the findings and recommendations below indicate, electronic records represent both a substantial challenge and a significant opportunity for NARA.

The committee's findings and recommendations are organized under and support the following six high-level recommendations:

1. Get ready for a rapidly rising tide of electronic records.
2. Plan for continuing technology change and increasing user expectations.
3. Reengineer relations with federal agencies to help them create records that are archive-ready.
4. Do not assume that ERA is unique: become more involved with other organizations that have interests in preserving electronic records.
5. Learn how to exploit the enthusiasm and capabilities of the research community and work with others who do that well.
6. Take strong measures internally and provide government-wide leadership to ensure record integrity and provenance.

These recommendations are discussed in more detail below.

RECOMMENDATION 1. PLAN FOR A RISING TIDE OF ELECTRONIC RECORDS

Finding 1.1 Within a short time, the vast majority of records will be electronic. The volume of records and the challenge of preserving them will continue to grow significantly.

Four technology effects are creating an avalanche of digital materials: the growing fraction of information that is born digital and not systematically retained on paper, the relative ease and inexpensiveness of recording information digitally compared with the effort and cost of recording it on paper, the advent of new technologies for communicating and recording information, and the decreasing cost of storage, which drives the retention of information that previously would have been discarded.

It is reasonable to expect that the growth in the amount of digital information produced in society at large will likewise occur in government—and thus the growth in the volume of permanent records. The volume of data to be stored will also grow because individual records

will continue to become larger. In addition, the total number and variety of record data types[3] that must be stored by NARA and other repositories will increase over time as new versions of existing types are created and entirely new ones are invented.

Recommendation 1.1 The National Archives and Records Administration (NARA) should devote a significant and growing fraction of its attention and resources to electronic records, especially in the areas of records management, appraisal, and transfer of records to NARA.

A rapid shift to electronic records has two important implications. First, as the findings and recommendations below detail, many of the processes associated with appraisal, scheduling, ingest (i.e., intake of records and associated metadata), and access will require major modification in order to accommodate the volume of electronic records expected. Second, as the share of records that are electronic continues to grow rapidly, the relative resources allocated for their management compared with resources for managing traditional records will need to grow as well. Over the coming decade or so, electronic records will become the primary line of business for NARA rather than a side activity. This change will have implications for NARA's organizational culture, structure, and staffing.

Finding 1.2 New forms of records will emerge as electronic systems capture new kinds of information that may also constitute essential evidence.

The way that government conducts its business has been transformed by the pervasive use of electronic systems. As those systems evolve, so too will the types of records that must be handled by NARA. Certainly, a great many formal government communications continue to use familiar, traditional records: forms, memoranda, pictures, and so on. But other types of information that could well be deemed "essential evidence"[4] are also being captured by electronic systems. They include new forms of communication, dynamically generated information, and transactions. Saving the potentially huge amounts of data associated with these new types of information, re-creating the behavior of old IT systems decades later, and establishing what should be retained as a permanent record are all extremely difficult challenges.

Recommendation 1.2 NARA should develop the capacity to predict and anticipate significant changes in record types and volume, styles of record keeping, and concepts of what should be preserved as a permanent record as well as the implications of such changes for future versions of the Electronic Records Archives (ERA).

Sustained growth in record volume, the continual introduction of new record types, new forms of record keeping, and new concepts of what constitutes a record mean that NARA will need the knowledge and agility to respond in a timely and effective manner so that it can make necessary changes to its systems and processes and so that it can advise and influence agencies as they implement new systems. To stay abreast of developments, NARA does not need detailed surveys, which are costly and difficult to conduct, but it does need to develop the capacity internally and externally (e.g., through research activities or advisory committees) to

[3] "Data types" is a more general term than "file formats," although the latter may be more familiar.

[4] "The mission of the National Archives and Records Administration is to ensure, for the Citizen and the Public Servant, for the President and the Congress and the Courts, ready access to essential evidence." National Archives and Records Administration (NARA). 2003. *Ready Access to Essential Evidence: The Strategic Plan of the National Archives and Records Administration.* NARA, Washington, D.C.

predict changes and discontinuities in record types by monitoring new data types, information production rates, and similar trends throughout society and within government.

RECOMMENDATION 2. PLAN FOR CONTINUING TECHNOLOGY CHANGE AND RISING USER EXPECTATIONS

Because plans at the outset are for the ERA to operate for a long time, considerable technology change will occur over the course of its existence. The committee does not have a crystal ball allowing it to see far into the technology future, but past experience does suggest some relevant technology trends that are likely to persist, as described below. Technology changes will, in turn, fuel user expectations: as new technology yields new capabilities that are deployed in systems outside NARA, users will expect the ERA to keep up. These expectations are also considered below.

Finding 2.1 Sustained performance improvements in the ERA's component technologies will continue indefinitely.

Although precise changes in the ERA's component technologies cannot be predicted, some general trends are apparent. For example, as in the past, new forms of storage providing such desired characteristics as higher volumetric density, lower cost, and faster access times will continue to emerge, and the commercial market will move toward these new technologies. Also, the familiar trends of ever-faster processors and networks will continue, making cost-effective network transfer and processing of ever-larger volumes of records possible. The past 20 years have seen sustained improvements in computing performance in many different respects, and there is no reason to suppose that such sustained improvements will not continue.

Recommendation 2.1 Because radical changes in information technology cannot be specifically anticipated, the ERA should be designed for change.

The continuous and significant change in technology capabilities into the indefinite future means that flexibility in the design of the Electronic Records Archives will be essential. The committee's first report stressed the need for a modular design for the ERA to allow components to evolve or to be replaced incrementally—advice that remains critical for the success of the ERA's initial design and future evolution. As the ERA evolves, improving its ability to change should be a major objective. For example, use of commercial off-the-shelf technology should be emphasized for the ERA, and ERA software should be written to be as portable as possible across offerings of different computer and storage vendors. Because network interfaces and protocols are likely to remain relatively stable, the ERA should conform to prevailing network standards. The ERA should also be designed to be readily extensible—that is, to accommodate change such as the introduction of new record formats without requiring immediate or widespread modification of the system.

Finding 2.2 The gains of new technologies will be highly skewed toward higher-volume products.

Certain computer technologies, such as ATA/IDE[5] disk drives and Ethernet networks, are high-volume, low-cost products for which the cost-performance ratio steadily improves. These

[5]The customary industry shorthand for a family of hard-drive interface standards (ATA, Advanced Technology Attachment; IDE, Integrated Drive Electronics); see also the subsection entitled "Exploiting the Commercial Mainstream" in Chapter 1.

products generally conform to de facto industry standards that have a relatively long lifetime, and backward compatibility is generally provided across several product generations. Low-volume products, in contrast, rarely exhibit sustainable cost-performance characteristics and often disappear rapidly from use. Today's trend in moving to disk from tape for mass storage is an example of what often happens in such cases, with a very high volume product replacing a relatively low volume product.

Recommendation 2.2 Components of the ERA should evolve in step with prevailing commercial trends.

The ERA will benefit from continuing technology evolution only if the ERA implementation stays roughly in step with mainstream commercial trends. NARA should therefore use high-volume products wherever possible. The functions of the ERA are not uncommon enough to support the use of unique hardware or software components; using such components would expose NARA to the risk of significantly greater costs and more difficult system evolution. NARA should be wary of special-purpose solutions that do not enjoy the benefits of continuous technology improvements driven by a large and active community of users. The ERA program should plan and budget for renovations of both hardware and software to stay on the technology curve of mainstream IT.

Finding 2.3 Full-content search, information retrieval, and information extraction offer inexpensive and reasonable, albeit imperfect, capabilities today for textual materials.

One of the major challenges for the ERA system is that of ingesting the anticipated volume of records without being bogged down by the human effort associated with producing the purely descriptive metadata that have historically been important for finding records. Realistically, there are unlikely to be enough resources (people or dollars) to rely on anything other than computer-based algorithms for this purpose.

One approach to reducing this burden is full-content searching. There has been considerable progress in this area—that an enormous body of publicly accessible Web pages and documents of multiple data types can be searched effectively by search engines such as Google is a major achievement of relevance to NARA. Furthermore, there is widespread demand for better searching technologies.

Another approach is to use information-retrieval techniques such as the family of techniques for automatic metadata extraction. These methods, which make use of information-extraction techniques to extract metadata from the content of records, have reached a level of maturity that permits them to be used in production systems; the techniques will continue to improve steadily, but no breakthroughs are on the horizon. Information-extraction techniques do not and will not provide the accuracy of a human. Nevertheless, they are much less expensive than manual processing is, and in some instances the margin of error they provide may be acceptable to NARA (and to NARA's customers, who may be more interested in having rapid access to a larger universe of archived records than in high precision). For the foreseeable future, automatic metadata extraction will need to be combined with expert guidance to achieve usable results.

Importantly, as is true with the provision of search capabilities, automatic metadata extraction need not be performed at ingest of records. The highest priority at ingest is capturing the records along with metadata that are not discernible from the content; the extraction of additional metadata can be deferred and can of course be repeatedly applied as techniques improve.

Finding 2.4 There is no general technical solution for the digital preservation problem today, and one is unlikely to appear any time soon.

As discussed in the committee's first report, the ability to create, store, and retrieve the bits that represent a record is a fundamental requirement of the ERA. Good implementations of this capability will be quite complex and will require refinement and evolution over time. However, the basic requirements for this part of the ERA are fairly well understood, and technologies exist or are being developed to satisfy most of them.

Ensuring that bits can be *understood* by systems and by human beings in the distant future is much more difficult, however. The committee's first report, citing the imperative of making significant progress toward implementing an archival system, proposed a pragmatic, short-term strategy. But as development of the ERA proceeds, NARA will need to seek out longer-term solutions.

For the long term, no generally applicable technical solution is evident. No single proposed approach to digital preservation, including migration, migration to durable fixed formats, or emulation, will work in all cases. Each method has its advantages; each has drawbacks and associated costs. One cannot know with any great certainty which proposed methods for preserving the interpretability of digital objects into the distant future will subsequently be deemed successful, much less relatively cost-effective. Preservation will therefore require a combination of approaches. Selecting the appropriate preservation techniques is part of the expertise that will be required of electronic records archivists. There will be a need for compromises based on user requirements and varying levels of service that NARA provides for different records.

The preservation problem is unlikely ever to be completely solved. Instead, our understanding will grow over time in IT areas that become relatively stable, such as relational databases and geographic information systems, and as digital preservation increasingly becomes a mainstream activity. In spite of this growth in understanding, new problems will arise as novel record data types emerge.

There is, however, a significant opportunity for NARA to ease some preservation problems by helping promote the development and adoption of formats that have desired attributes for preservation. These attributes include public specifications, broad applicability, compatibility with prior-format versions, and stability. NARA has already been involved in the development of one such format, the nascent Portable Document Format (PDF)/A standard. NARA will need to determine and periodically amend the list of formats that it preferentially supports and work to ensure that record-creating agencies provide records in these supported formats. This is an important part of ensuring that records are created archive-ready (see Recommendation 3.1).

Finding 2.5 User expectations will be high and will continue to grow as the capabilities offered by other information technology (IT) services improve.

Users are likely to expect NARA to keep pace with the aggressive rate of improvements that they see in commercial online services. Also, it is reasonable to expect that users will want appropriate transformation and processing—such as conversion from persistent forms into easily usable forms—and not just access to raw data. In other words, users will expect services well beyond the simple delivery of static records.

Given its limited resources, NARA cannot afford to meet all of these expectations for all

records. Expectations will need to be reconciled with the notion of varying levels of service for different record types. One such adjustment would be to make all records easily accessible even though they were not all equally easy to view or process. Another approach would be to enable third parties to meet the requirements of users who have special retrieval or processing needs or who must work with infrequently used document formats (see Recommendation 4.1).

RECOMMENDATION 3. REENGINEER PROCESSES TO ACCOMMODATE CHALLENGES AND OPPORTUNITIES PRESENTED BY ELECTRONIC RECORDS

If NARA and federal agencies do not succeed in transforming their approach to transferring permanent records for preservation, the probable result will be significant difficulty in acquiring the desired permanent records and a costly, labor-intensive effort applied to those records that are acquired. As the subrecommendations below detail, NARA should therefore establish procedures and systems to ingest new records automatically on a routine basis, capture metadata at or close to record creation, take steps to increase the likelihood that records are provided archive-ready, and use automatic techniques to supply missing metadata. In implementing the recommendations below, NARA should determine the areas in which existing law and regulations provide sufficient flexibility and those in which changes to the policies governing records management may be required.

Finding 3.1 The growing volume of electronic records will swamp a system that relies on manual processing.

As the volume of electronic records to be stored in the ERA grows (Recommendation 1), existing processes, which rely on manual handling of records, will cease to be cost-effective, or even feasible. Affordability requires automation. It is, for example, not feasible to manually supply additional metadata about ingested records except at a very coarse level of granularity.

Finding 3.2 Preparation for the ingest of electronic records closer to their creation will make archiving easier, cheaper, and more accurate.

The ingest of electronic records decades after their creation could be problematic. This is because obsolete data and metadata formats may be difficult to interpret, and it may be quite difficult to modify the systems in which the records reside in order to facilitate ingest.

There are two very similar techniques for addressing this problem: a passive one, in which NARA is sent records early, on a provisional basis, and carries out a "dry run" of ingest; and an active one, in which NARA actually ingests records in anticipation of a future transfer of custody to NARA. Both give early warning of potential problems with ingest. Although the active mode gives NARA the most information about the records that it will have to handle, it is probably harder for agencies (and NARA) to implement. On the other hand, the resulting archival copies might also provide additional incentives for agencies to support the creation of archive-ready records and their delivery to NARA, as the ERA could provide off-site backup of agency records.

Finding 3.3 Automated ingest depends on records being provided in structured, standardized formats together with sufficiently standardized essential metadata.

Without some degree of standardization in the essential metadata (metadata that cannot be derived from the record itself through information extraction or search), the ingest of each

set of records will require significant, costly human intervention. Nonetheless, NARA and the ERA system will have to be flexible in accepting records from originating agencies—that is, neither NARA processes nor the ERA design should be brittle with respect to variations resulting from legacy systems, poor implementation of unusual record types, and other circumstances.

Recommendation 3.1 To permit records to be ingested using largely automated processes, NARA should strive to change federal agency practice and system design so that permanent electronic records are created or provided to NARA "archive-ready."

In order to achieve the requisite degree of standardization of formats and essential metadata, NARA will need to provide very explicit guidance and realistic requirements that record-keeping software systems can support. These would need to be requirements that agencies can implement without significant changes to normal business processes or without investing in large amounts of human effort to provide needed metadata. In particular, NARA should consider requiring all newly acquired agency systems that produce permanent records to do the following: create those records in formats acceptable to NARA, include explicit metadata in their output, and use standardized mechanisms for transferring records to NARA, such as using secure network communications. In the long run, archiving considerations will have to become part of the government process of software procurement and development for systems that produce or are deemed likely to produce permanent records.

When a government agency does indeed revise or reengineer its business practices, it can (and sometimes does) vastly improve its records-management practices. These reengineering efforts are opportunities for NARA to furnish the guidance and standards that ensure that archive-ready records are created by the reengineered system.

Successful implementation of the ERA system may also require NARA to become more actively involved in the establishment of standards used in constructing federal systems, such as those governing the formats used to represent records and associated metadata. It may also require NARA—or other bodies responsible for information policy, such as the Office of Management and Budget—to establish guidance to agencies, together with auditing and enforcement tools as necessary.

Experience shows the difficulty of implementing systems that depend on user-supplied metadata or that depend on user compliance with externally imposed requirements that take effort and do not necessarily support the user's internal business needs. Consequently, NARA should foster the development of software that automatically captures metadata at record creation. It should also work closely with agencies to determine whether its requirements can realistically be implemented and link archive-ready concerns with agency interests in supporting the agencies' own record-keeping operations.

Finding 3.4 NARA should not let the existing backlog of electronic records that are not archive-ready undermine the long-term objectives of developing capabilities and processes to handle new records through automation.

Over the long term, the preceding recommendations address the move to automated ingest of records. However, a considerable backlog of electronic records already exists, and it will continue to grow before automatic measures take full effect. If NARA is to avoid being totally swamped, it will have to perform triage on the handling of the backlog. This will mean avoiding significant manual processing, preserving the obvious context and features of the

environment that would support extraction of automatic metadata in the future, and using automatic metadata extraction (and accepting its shortcomings).

Recommendation 3.2 The ERA system should include fallback capabilities that do not require intensive manual processing for instances in which agencies do not provide records archive-ready. NARA should accept and accommodate the resulting varying levels of metadata quality.

There will be many cases, including those involving legacy and noncompliant systems, in which the specified or desired metadata are not provided along with agency records. As a fallback, NARA should employ low-cost if imperfect techniques for automatic metadata extraction. Such techniques are a much more cost-effective way than is manual processing for handling enormous volumes of records. Using these techniques will require NARA to accept significant imperfection in some of its metadata and thus to shift from the current mindset of having very high quality metadata.

It will also be necessary for ERA systems and processes to carefully track the provenance of metadata and to distinguish among different metadata sources, which have varying quality. As metadata-extraction techniques improve, the metadata can be upgraded; the ERA should accommodate easily making such updates to record descriptors. Improved search techniques can also be used to improve future access.

Recommendation 3.3 NARA should carefully consider its approach to the description, appraisal, and selection of electronic records for retention. The trade-off between cost and utility should always be considered.

As discussed above, it is essential for the ERA to able to ingest records automatically and to have the agencies that create records take on greater responsibility for providing required metadata. However, this does not mean that agencies would have to take on the full burden traditionally borne by archivists.

The costs and benefits for varying levels of investment in metadata (using either manual or automatic processing) should be considered carefully. Electronic records present significant opportunities to exploit new capabilities, such as full-content searching, structured documents, and automated metadata creation. To the extent that NARA can make use of full-content searching, and exploit search and other access mechanisms that agencies themselves have created for their own records, there will be opportunities to create new kinds of finding aids that do not rely on manual processing.

Also, there are trade-offs between what is gained through fine-level appraisal versus the costs of making the selection of electronic records for retention. There is a growing volume of information being produced, there are ever-improving automated techniques for finding records, and a rapidly falling cost of storage. Thus, instead of examining records carefully to mark only a select few for retention, it may become more expedient, and more appropriate, to retain larger "chunks" of records (even if some of the records may not warrant retention as permanent records), in order to ensure cost-effective preservation of the government's output. Such a shift would appear to be consistent with NARA's Records Management Initiative, with its emphasis on functional appraisal, which implicitly is more broadly granular.

Recommendation 3.4 NARA should continue to expand the use of cooperative arrangements that shift responsibility for preserving and providing access to records to the agencies that create them.

The model of having certain agencies retain responsibility for certain types of records by operating what are, in effect, affiliated archives, is well established (e.g., the arrangements covering the preservation of large scientific data sets; a NARA-Government Printing Office agreement). This model could be expanded. Under such arrangements, NARA's role shifts to that of a standards setter, a records-access federator, and a repository of last resort.

NARA would benefit from allowing more agencies to operate affiliated archives because it could then spend its scarce resources on problems that the agencies cannot address on their own. Such arrangements would also provide a way for NARA to partner with agencies to provide better access to records. Advanced access services are best provided or supported by the mission agency that knows the domain; in this way the requisite expertise in managing a complex system need not be transferred to NARA. Advanced access capabilities can also be provided by third parties, possibly for a fee (see Recommendation 4.1). With these opportunities comes the responsibility for establishing and verifying appropriate standards for affiliated archives and for establishing fallback mechanisms in the event that agencies cannot fulfill their responsibilities at some point in the future.

For such partnerships to work, NARA will have to offer some form of incentive. For example, as NARA develops its electronic capabilities, it has the opportunity to provide generic expertise in electronic records management and preservation that complements the subject-matter expertise of the partner agencies. Finally, as the number of partnerships grows, NARA can collect, codify, and disseminate best preservation practices.

In addition to depending on NARA's contributions, the success of cooperative agreements will depend on agencies honoring their commitments under these agreements. NARA may need to assume the role of auditor.

RECOMMENDATION 4. ENABLE PARTNERSHIPS WITH OTHER ORGANIZATIONS WITH INTERESTS IN ELECTRONIC RECORD PRESERVATION

In the long run, the ERA should become part of a larger collection of archives, libraries, and information services operated by governments, nonprofit organizations, and the private sector rather than a unique, stand-alone system that only NARA runs. Many other organizations are also developing large collections of electronic materials and associated systems and tools—notably, initiatives launched by the Library of Congress, the Government Printing Office, and the National Library of Medicine.

Finding 4.1 An open architecture with public interfaces would allow NARA to outsource some functionality, encourage competition of services, and foster innovation in access.

By designing the ERA system so that third parties can build new capabilities on top of it, NARA can accomplish several things. It can, at low cost, tap the resources and entrepreneurship of other organizations, commercial firms, and researchers to provide enhanced services, help fulfill NARA's mission, increase the perceived value of NARA's holdings, and explore new technologies and approaches.

Recommendation 4.1 NARA should use a design strategy for the ERA that enables a range of other organizations to provide access services that NARA does not offer.

NARA does not have to be the sole access provider. Instead, it can enable agency partners and private service providers to develop alternative access mechanisms. Different providers

could, for example, offer different views of the same underlying data, index them differently, or derive new digital products from them. Capabilities that might be built on top of the ERA system include the following: data annotation to supplement and correct record metadata through third-party systems that are separate from NARA, but accessible as if part of a unified system; the continued introduction of specialized, enhanced capabilities by partners and vendors that support particular needs; and entrepreneurship to provide value-added services, either through application programming interfaces (APIs) or bulk download.

The result—and the desired outcome—is that NARA would be able to leverage considerable assistance from its customers, partners, and other third parties. Once the technical capabilities were in place, NARA could take the concept one step farther by actively seeking out partners to work in areas of interest. This effort would provide NARA with a low-cost method of exploring new technologies and staying on the leading edge.

Recommendation 4.2 NARA should anticipate future demand for federated access to records in the ERA and other affliated archives.

Federated access, in which a software layer allows one to access a collection of archives as if it were a single archive, would be a valuable capability. Federation would provide a common interface to federal records regardless of which agency happened to have responsibility for managing them. (Recommendation 3.4, above, recommends distributing some of the burden for archiving to institutions beyond NARA.) Federated access would also be valuable in allowing access to span NARA and presidential library resources, for example. Early iterations of the ERA design may well not include these capabilities, but the ERA should be designed with their future addition in mind.

Although there is a large relevant technology and standards base, there is currently no real corresponding base of operational practice in federating access to archives. Among the issues to consider are the minimum common standards that archives must agree to (e.g., unique record identifiers) to support federated access. The easiest way to accommodate federation might be to expose an API that accommodates it. And if that API is exposed, federation could be provided by third parties with minimal effort by NARA.

RECOMMENDATION 5. BROADEN INTERACTIONS WITH THE RESEARCH COMMUNITY

Finding 5.1 NARA has been too narrowly engaged with the research and development occurring in digital archiving, digital libraries, and related areas.

In the early stages of the ERA program, NARA sponsored several research and development projects in which it worked with limited segments of the relevant research community consisting of a small number of specific organizations. It has not, however, fully exploited opportunities to engage the broader research communities working in areas related to NARA's interests. Prototypes are a useful part of the system development process and a valuable way to test and communicate the results of research, but prototypes by themselves do not advance the state of the art or science.

Recommendation 5.1 NARA should be more actively engaged with multiple and diverse sources of research so as to gain early experience with emerging technologies, build internal expertise, and ensure that its future needs are met.

By virtue of its extensive experience in archiving and its significant needs for preserving electronic materials, NARA should more actively engage relevant research communities—for example, by sponsoring research, suggesting problems that need to be addressed, engaging with researchers, hosting workshops and seminars, and hiring students. These activities would help meet NARA's needs and those of the broader archival and preservation communities.

An organization like NARA should be involved in research for several reasons: (1) to gain early experience with technologies that might be expected to begin appearing in commercial products within roughly 3 to 5 years, (2) to learn more broadly about a research area through direct and regular engagement with related research communities and the experts within those communities, (3) to describe to researchers challenging problems faced by NARA, and (4) to build absorptive capacity to understand how to apply the results of relevant research sponsored by other organizations.

NARA should anticipate that since research involves a degree of risk, not all research projects will necessarily produce results that can be implemented directly. The trade-off is that NARA stands to benefit from innovative ideas that had not been anticipated.

To manage its research activities most effectively, NARA should seek to employ individuals who have existing relationships with the various relevant research communities. Such individuals should also bring with them a proven track record of producing or applying research results in these areas.

Finding 5.2 Most of the technical challenges faced by NARA are shared with other organizations operating large repositories for digital materials, but a few problems are specific to the operation of NARA and similar archives.

Two distinct classes of research areas in which NARA appears to have unmet technology needs are apparent:

1. *Problems shared with other organizations.* Most of NARA's technical problems are the same as those faced by designers and operators of any large digital library or repository. For these types of problems, NARA's greatest leverage will come from drawing on research sponsored by other organizations, learning about best practices, inducing others to work on problems of interest to NARA by offering corpora, and participating in joint research programs with other federal agencies and organizations that face similar challenges. Wherever technologies may be shared with other applications, partnerships to jointly address problems involving these technologies will help stretch limited research resources. By joining with others, NARA also can couple to a joint learning curve—that is, it won't have to make unique mistakes.

2. *Problems specific to government archives and similar institutions.* Examples of research problems of particular importance to NARA and similar institutions include the following: how government transactions systems should be preserved, how to provide digital assurances over a very long period of time, and how to provide semiautomated redaction and declassification of materials. Working on these more specialized problems may require engagement with specific existing research communities. Partnership with agencies that have greater experience in managing IT research programs is also likely to be the most effective mechanism for this class of research problems.

Recommendation 5.2 NARA should develop a research strategy that lays out its technical needs and the partnerships that would best address them.

Considering NARA's limited resources, its limited capacity to manage research programs, and a set of challenging technical problems related to the general topic of archiving and digital preservation, the issues of how and where NARA should invest arise. Questions for NARA to consider include these: What agencies share an interest in NARA's problem or a similar problem? What research community is working on that problem or a similar problem? How can NARA most effectively join and influence that community? How can it learn from or participate in others' research programs? What results are anticipated, and how will the results be brought back to NARA?

The development and deployment of the ERA will engender a number of research topics. Some of these topics can perhaps be anticipated now, but others will arise only as experience with the ERA accrues. Having a research agenda will help guide NARA in selecting which research initiatives and organizations to become involved with. As NARA defines its research agenda, it should be careful to frame its needs as problems that academic and commercial researchers can understand and address, and it should frame them in ways that are amenable to working in partnership with other organizations.

Developing a research agenda is likely to be daunting when an organization is in the midst of a rapid technological and cultural shift, as is NARA today. Fortunately, NARA need not (and should not) work in isolation or start from scratch in developing a research agenda. Partnering with more experienced research organizations, as recommended below (see Recommendation 5.3), will help, but NARA will nonetheless need to develop its own voice. Several recent research agendas in the area of archiving and preservation (including reports prepared for the National Science Foundation [NSF], the Library of Congress, and NARA's National Historical Publications and Records Commission) provide a basis for NARA to find common ground with other research initiatives.

Recommendation 5.3 NARA should implement its research strategy by working with other agencies that have similar research needs and expertise in managing research programs.

Implementation of NARA's research strategy will require parsing the agenda into a research program, building or joining a community of researchers, monitoring work, and creating incentives and opportunities for technology transfer. In carrying out all of these activities, NARA should seek partnerships with agencies for which research management is a core competency, such as NSF or the Defense Advanced Research Projects Agency (DARPA). These agencies have developed considerable expertise with selecting and reviewing research, engaging with research communities, and managing research projects. The selection of specific research proposals and routine management of the research process should be left to organizations for which research management is a core competence. NARA should work with its partners to ensure that it has a voice in developing research initiatives and has opportunities to engage directly with the organizations and experts that do the research.

NARA's fundamental requirements for digital preservation and access are shared to varying degrees with a number of government organizations that also must maintain access to digital records into the indefinite future. Problems such as semiautomatic redaction or declassification can be investigated with partners such as DARPA or the intelligence community's Advanced Research and Development Activity (ARDA). Another opportunity is presented by

NSF's program in digital preservation, which is explicitly designed to support multiagency participation.

Recommendation 5.4 NARA should also engage the research community by providing access to data in the form of electronic records.

Academic researchers are often starved for data: they have good ideas, but no way to test them in the real world. Providing interesting data in a form that the research community can use is often a sufficient incentive to get the best people working on a problem. It also provides a very cost-effective way of engaging a research community. The National Institute of Standards and Technology and a number of university groups have experience in this area; partnering with such an organization would help NARA learn how to create data sets that are interesting and useful to various research communities.

NARA might also consider providing direct access to some of its digital archives in the form of APIs that provide access to selected internal data or operations of the ERA system, enabling people to build new interfaces or applications on top of NARA's archives. Such access would be attractive to a variety of computer science and social science research communities, and it could be provided by NARA at relatively low cost, considering the potential benefits.

RECOMMENDATION 6. ENHANCE CAPABILITIES AND PROVIDE GOVERNMENT-WIDE LEADERSHIP FOR RECORD INTEGRITY AND PROVENANCE

As NARA recognizes, electronic records are subject to a variety of threats, ranging from accidental corruption to deliberate tampering, that are quite different in character from the threats to which paper records are subject. It is essential for the ERA and the records-management practices that it supports or requires to be up to the task of adequately protecting the public record.

Fortunately, electronic records also afford important opportunities to improve integrity and authenticity assurance through the use of robust cryptographic techniques, automatic replication, and system design. In order to ensure trust in the digital records held by the ERA, stringent measures—including the use of cryptographic techniques—should be taken to protect records against deliberate or accidental compromise. These methods should be built into the ERA system design and its operational procedures. The overall trustworthiness of records also depends on the methods used by agencies that supply records to the ERA. Thus, NARA should concern itself with the trustworthiness of a record throughout its entire custody chain, and it should provide leadership on these issues.

Finding 6.1 NARA's past procedures and practices would be inadequate for safeguarding digital records.

Concepts developed and applied to paper records, such as chain of custody, are also of value in assessing the authenticity of electronic records. They do not, however, adequately address the threats associated with electronic records. These threats include hardware, software, and operational errors, as well as individuals (both outsiders and insiders) who may either tamper in subtle ways with individual records or systematically modify large numbers of records.

Recommendation 6.1 The ERA should use appropriate technical and procedural methods available for assuring the integrity and authenticity of electronic records.

The new vulnerabilities introduced by virtue of the digital nature of the records to be preserved underscore the importance of complementing traditional procedures by making use of robust digital techniques and an appropriate overall system design for verifying integrity and authenticity.

Digital assurances for records are based fundamentally on maintaining *multiple, geographically and administratively separated copies* and on two cryptographic tools—*hash digests for integrity checking* and *digital signatures for authentication.* Digital signatures are an excellent technique for verifying that recently transmitted data—such as a set of records being transferred from an agency to NARA—actually came from where they were supposed to have come, but digital signatures are a poor means of verifying the origin of stored data. Instead, the origin can be verified (1) at ingest, by recording metadata on the outcome of having verified the origin of those data at the time they were received, and (2) from then on, by maintaining the integrity of the stored data and associated metadata.

Designing digital assurances into an electronic records archive is similar to designing security measures: that is, the cryptographic techniques must be chosen carefully, but more importantly, the overall system design, not just the cryptographic mechanisms, must not allow openings for attackers. And as with security systems, provisions must be made to change the cryptographic algorithms or system design if the chosen cryptographic algorithms are ever found to be faulty (a very likely contingency at some time in the future) or if the system design is found to be susceptible to attack.

Cryptographic protections are only a part of the solution, however, in part because they detect damage or attack only after the fact. Prevention and repair are even more crucial parts of the solution and depend on careful procedures and system designs, including the following: (1) maintaining multiple copies that are geographically separated and independently administered; (2) tracking records from their origins, through storage and IT systems of various sorts, and eventually into the archive; (3) protecting the systems that hold records from unauthorized tampering, both by internal employees and external attackers; (4) employing access controls, audit logs, and procedural safeguards when changing the archive contents to reduce human errors (e.g., requiring multiple people to authorize a change); and (5) employing sound software development methodologies to reduce the likelihood of record corruption due to software bugs.

Recommendation 6.2 Throughout the ERA system design and operation, NARA should carry out an iterative threat modeling, analysis, and response process.

A threat analysis should be done early in the design of the ERA so that proper security measures can be designed and implemented from the outset. It is essential that ERA design proposals be analyzed against a threat model in order to gain an understanding of the degree to which alternative designs are vulnerable to attack. The threat analysis should be undertaken not only for the ERA itself, but also for the entire chain of custody of a record from its creation, through its retention for active use in a government agency, to its eventual transfer to the archive. This initial threat modeling would be only the first step of a larger, iterative threat-countering process that involved designing against expected threats, observing failures that occur, and designing new countermeasures. Threat modeling and analysis are a specialized

art, which suggests that NARA should seek outside specialists to undertake these tasks under contract.

Recommendation 6.3 The ERA should be designed so that auditing, monitoring, and testing of ongoing operations verify that digital assurance measures are working as intended.

Methods for detecting very rare events such as the corruption of stored information are extremely susceptible to error because they are rarely exercised. The ERA should be designed to include ongoing processes that detect errors and that exercise error-detection and correction procedures. One of the threats to ERA integrity is software errors that corrupt file structures. For this reason, procedures associated with introducing new software must provide for adequate testing, gradual introduction of software changes so that impacts on data integrity are apparent early on, and capabilities to repair any errors that are introduced.

Recommendation 6.4 NARA should ensure that it has access to expertise in digital assurance techniques and in the operational procedures that accompany them.

The basic cryptographic techniques and other trustworthiness measures discussed in this recommendation are widely used in a variety of contexts. However, their application in a system designed to provide very long term assurances of authenticity, integrity, and chain-of-custody management is likely to be leading edge for the foreseeable future. As a result, NARA will need access to highly specialized expertise on an ongoing basis—for example, to assess initial and subsequent system designs and operational procedures. As it would seem difficult for NARA to maintain this kind of highly specialized expertise in-house, it should consider establishing an expert advisory committee to specifically address the kinds of issues discussed in this recommendation. Digital assurance in the context of long-term preservation is also a possible research area for NARA.

Recommendation 6.5 NARA should lead an effort to enhance digital assurance for electronic records government-wide.

Agencies also need to protect records, lest NARA's archiving activities turn into a "garbage-in, garbage-out" exercise. All of the subfindings and subrecommendations presented above are equally applicable to the agencies that originate records and later pass them to NARA and to agencies that manage records locally under cooperative agreements.

NARA's role in records management offers an opportunity to establish federal standards and to inaugurate, advance, and even mandate their use in records management. By virtue of its central role and planned longevity, the ERA is the ultimate client, and perhaps the one with the most stringent requirements, for secure digital records management. NARA has undertaken leadership in other areas of records management through the Records Management Redesign program and the E-government Electronic Records Management initiative and has issued records-management guidance in several areas related to digital assurance. NARA should help spread the use of digital assurance techniques, as well as monitor their use in records management, throughout the government.

CONCLUDING REMARKS

As the National Archives and Records Administration recognizes, its initiatives for electronic records are critical to fulfilling its mission to preserve essential evidence as society and

government information continues to move to a born-digital and even only-digital state. In fulfilling its mission, NARA enters an era in which it must understand, cope with, and embrace the rapid pace of change that characterizes information technology. As these changes unfold, NARA will face a number of difficult questions, some of them alluded to above.

The rapid evolution in government information and information systems provides a good illustration. Formal government communications continue to be conventional records: forms, letters, pictures, and so on. But citizens and government officials increasingly interact via e-mail and text and voice messages, not memoranda and forms. Web pages increasingly present not only static documents but also ephemeral views that are not stored as records, but are merely computed from online databases through software that itself changes over time. What portions of this constantly changing material are essential evidence that documents the national experience? Handling the resulting flood of data and re-creating the behavior of ever-more-complex IT systems decades later are both extremely difficult challenges. The National Records and Archives Administration of an e-government will have to be equipped to address the technical challenges that arise, and to examine and possibly revise how it carries out its basic mission.

1

Ongoing Technology Change and Rising User Expectations

AN AVALANCHE OF DIGITAL INFORMATION

Several important trends are driving the increase in the volume of digital records: more information is being produced and stored, a greater fraction of information is born digital, and a greater fraction of the born-digital material is retained only in digital form.

A study conducted at the University of California, Berkeley, estimated at about 30 percent a year the rate of growth in the total amount of new information stored between 1999 and 2002.[1] Much of this information is being produced and stored in digital form. Although it is difficult to isolate the amount of federal government information in these general data, there is no reason to assume that the federal government is not experiencing these broader trends.

A growing percentage of information is born digital, stored in file servers, database systems, correspondence-management systems, and other electronic information systems, and not systematically retained on paper. The University of California, Berkeley, study found that the amount of information printed on paper is still increasing but that printed documents represent less than 1 percent of the total information being produced today and that the vast majority of information is being stored on magnetic media. This result seems reasonable, given anecdotal evidence that the vast majority of documents, for example, are created using office automation software. No paper copy may ever be produced of many of the various born-digital records that are stored in file servers, database systems, correspondence-management systems, and other electronic information systems. As electronic filing systems become more commonplace, it will become less likely than in the past for the paper copies that are printed to be retained systematically.

[1]Peter Lyman and Hal R. Varian. 2003. *How Much Information, 2003*. School of Information Management and Systems, University of California, Berkeley. Available online at <http://www.sims.berkeley.edu/how-much-info-2003>. Accessed January 7, 2004.

This trend away from paper documents is illustrated by the experience of the Government Printing Office, which reports that the number of documents that it prints has fallen by roughly two-thirds over the past decade. People are increasingly accessing documents via the Web, and many publications are no longer produced in paper form at all.

Five major technology trends support a prediction that the National Archives and Records Administration (NARA) will soon face an avalanche of digital materials:

- *Computer technology makes it easier and less expensive to record more information than can be recorded in the paper world.* The increasing use of automated systems will make it possible to capture a larger number of records associated with government functions. Automated e-government services and systems, for example, will enable and encourage more transactions per citizen and per government employee.
- *Information and transactions are finer-grained*. Information is becoming "finer-grained." For example, it will be possible to record—and preserve—all of the individual changes made to a case file as well as its final contents. Content/document-management systems allow an audit trail for every individual who "touches" (i.e., creates, displays, copies, or modifies) a record. This audit trail metadata could be quite voluminous. Also, new kinds of records will be created as data from government-operated sensors are automatically logged.

Many record-keeping systems are transaction-based, with databases whose contents are continually changing. In many cases, it may be impractical to log and preserve all of these transactions. The alternative is that systems are designed so that snapshots are captured and preserved on a regular basis. This way of preserving data will require changes in system design and careful consideration about what should be captured, what format it should be captured in, what the refresh rate should be, and so forth.

- *Information is often dynamically generated.* Information used by the public and by decision makers is increasingly created on-the-fly. For example, many Web pages today are dynamically created and formatted for presentation from underlying databases. The problem of archiving databases that change over time is not a new one, but archivists will face more complex preservation decisions as more and more information is dynamically generated. Even when the underlying database itself is amenable to preservation, displaying information from that database for users poses enormous challenges. For example, the information displayed might be the output of analysis software that has gone through multiple revisions—what was actually viewed on any particular day is a function of the version of the software running on that day. Capturing such changes and transformations is both challenging and crucial.
- *Technologies will allow people to communicate and record information in new ways.* In the 1990s, some e-mail messages began to be determined to be permanent records. Likewise, technologies today are creating records that may be selected in the future for permanent preservation; examples include video conferencing, instant messaging, voice over Internet Protocol, and video presentations.
- *The drop in storage costs drives the retention of information that previously would have been discarded.* The cost per unit of storage currently falls by roughly a factor of two per year.[2] As a

[2]For a detailed examination of trends in disk drive storage, see, for example, H. Grochowski and R.D. Halem, 2003, "Technological Impact of Magnetic Hard Disk Drives on Storage Systems," *IBM Systems Journal* 42(2).

result, enormous quantities of information are being stored today. The University of California, Berkeley, study estimates, for example, that just over 400,000 terabytes of original information are stored each year on hard disks.[3] Indeed, storage costs have dropped so much in the past decade that many individuals and organizations have stopped "garbage collecting" (i.e., deleting old files) within their stored information. As this trend continues, businesses will increasingly destroy electronic records on the basis of the business risks of retaining records beyond regulatory mandates rather than because of storage costs. With digital storage growing ever cheaper, it becomes more economical to keep everything than it is to put in the effort needed to decide what can be thrown away. Tools capable of searching ever-larger volumes of information will help fuel demand for more things to be considered permanent records.

The volume of data to be stored will also grow because individual records will continue to become larger. It is becoming easier and more commonplace to generate and digitally store multimedia and other rich content. A 30-minute video file is much larger than a 30-minute audio file. A PowerPoint presentation with embedded animation is much larger than a plaintext file.

Finally, the total number and variety of record data types that must be stored by the National Archives and Records Administration and other repositories will increase over time as new versions of existing data types and entirely new data types are invented.[4]

The overall growth in the variety of data types will be accompanied by spurts of growth and contraction in the number of data types in active use at any point in time: as a new class of record types emerges, one will likely see an initial explosion of data types. That proliferation will occur as various vendors introduce products, and it will be followed by a consolidation to a smaller number of de facto standards as the market picks winners and losers. Over time, barriers to market entry will tend to increase as an industry segment matures; those barriers will lead to a smaller number of vendors and higher expectations of greater interchangeability. There were, for example, many early word processors, each with its own file format, but today

[3]Peter Lyman and Hal R. Varian. 2003. *How Much Information, 2003*. School of Information Management and Systems, University of California, Berkeley. Available online at <http://www.sims.berkeley.edu/how-much-info-2003>. Accessed January 7, 2004.

[4]A data type refers to the data-encoding rules describing how various kinds of records—word-processing documents, e-mail messages, images—are expressed as a collection of bits. For example, an image might be represented by bits whose data type is TIFF, GIF, or JPEG, where the specification describes how the bits are to be interpreted as an image. The more commonplace term "file format" is often used interchangeably with data type, but data type is more appropriate. The literal interpretation of file format is a specification for a file made up of bits. In some cases the bits might be embedded within a file (e.g., a file might contain a folder that contains multiple e-mail messages, one of which contains an image of data type GIF). Some data types are wrappers that encapsulate other data types. For example, files of data type WAVE may contain linear pulse code modulated or highly compressed audio. Other data types support bundling of multiple collections of bits that represent a single record together with associated metadata. Compounding the complexity of the problem are the myriad options and subtypes associated with some data types. For example, TIFF, which is a well-established data type for images, specifies a wrapper that can contain several different image encodings. Newer image data types such as JPEG-2000 are even more complex. The Library of Congress Web site ("Digital Formats for Library of Congress Collections," <http://www.digitalpreservation.gov/formats/>, accessed May 1, 2005), from which this discussion was partially drawn, provides information about various data types.

only a small number of data types are commonly used. Similarly, early Web sites depended on numerous ad hoc scripts, but today's Web sites are commonly implemented using more standard approaches. Today, the broad range of Extensible Markup Language (XML)-related formats is relatively immature and can be expected to follow a similar pattern.

Despite expected changes over time, for the foreseeable future there will be many data types and subtypes in use, creating great difficulty for archiving. NARA will never be able to support all of these types and subtypes equally well, yet important records may be created in these formats. Compromises in terms of the level of service provided will be necessary (see the section "Long-Term Preservation," below).

NARA will need to monitor the kinds of trends described above for several reasons. The volume and types of electronic records that it must handle have implications for the resources devoted to them. NARA must be able to anticipate trends in order to be able to make necessary changes to its own systems, to issue requests for proposals (RFPs) with sufficient scalability requirements, and to advise and influence agencies as they implement new systems. The objective is to avoid surprises like the one that occurred in the 1990s when use of e-mail became widespread and it proved difficult to adapt to this novel form of records.

It is neither essential nor cost-effective to carry out detailed surveys of potential types and volumes of records. Internally, NARA will need staff with the expertise and responsibility to track changes. External advice, such as that obtained through research activities or advisory committees, can supplement NARA's capacity to predict changes and discontinuities in record types and volume. Such external advice would be based on an awareness of new data types on the horizon and on an understanding of information production inside government and more broadly in society.

PLANNING FOR CONTINUED TECHNOLOGY CHANGE

It is envisioned that NARA's Electronic Records Archives (ERA) will operate for many decades. The enormous changes in information technology (IT) that have occurred just in the past 10 years suggest that the ERA will experience considerable technology change during its existence. These changes will come about in terms of both the types of records that the ERA will need to accommodate and the technology components that will be available for use in future iterations of the ERA. Indeed, the technology context and the system itself are likely to evolve during the ERA's lifetime—so much so that the problem can be usefully viewed largely in terms of how the ERA system will evolve around the archive's data and the data structures used to represent them, both of which are comparatively stable.

An important component of planning for the ERA's evolution will be to identify and distinguish various trends—those requiring system change, those requiring system additions, and those requiring simply that the system be sufficiently scalable.

Technology Performance Improvements

Although the committee does not have a crystal ball for seeing far into the technology future, past experience suggests some relevant technology trends that are likely to persist. Precise changes in the ERA's component technologies, for example, cannot be predicted, but some general trends are apparent. Storage will continue to become much denser and cheaper. As in the past, new forms of storage providing such desired characteristics as higher volumet-

ric density, lower cost, and faster access times will continue to emerge, and the commercial market will move toward these new technologies. As was discussed in the committee's first report, at present, tape-based storage is gradually being replaced by magnetic disk storage. Also, the familiar trend of processors and networks becoming ever faster will continue, making possible cost-effective network transfer and processing of ever-larger volumes of records.

The past 20 years have seen dramatic performance improvements in many directions. In 1984, for example, a state-of-the-art workstation had a processor that operated at 1 million instructions per second and had 1 megabyte of RAM. In 2004, a competent off-the-shelf personal computer (PC), which cost much less than the 1984 workstation, had a processor with a clock rate of around 2 gigahertz and 1 gigabyte of RAM. In 1984 a 10 megabyte hard drive for a PC retailed for $2,500; today a 250 gigabyte hard drive for a PC retails for under $200. The capacity is up by a factor of 25,000, and the cost per byte is down by a factor of 300,000. In 1984, the fastest links in the Internet were 64 kilobits per second; today the fastest links are 40 gigabits per second—up by a factor of nearly a million.

Experience has shown that it typically takes a decade or more to figure out what can be done with such new capabilities. For example, all of the Internet technology necessary to support the World Wide Web existed by 1982. That included a working Internet, a good-sized community of users with a need to communicate organized information of various types, the concept of hypertext, the Domain Name System, and desktop PCs connected to the network. But the World Wide Web itself wasn't invented until 1991—it took a full 9 years to work out the implications and realize that this application was possible. Therefore, even if technology improvements were to stop in their tracks tomorrow, one would still expect a very high rate of change (improvement in performance and cost) in delivered systems for at least another two decades. And there is no reason to suppose that these sorts of sustained performance improvements will not continue into the indefinite future.

Exploiting the Commercial Mainstream

The ERA system will benefit from continuing technology evolution only if the ERA implementation stays roughly in step with mainstream commercial trends. The functions of the ERA are not uncommon enough to support the use of unique hardware or software components, and relying on such components would expose NARA to the risk of significantly greater costs and more difficult system evolution. There will be, to be sure, some solutions that lead the way and appear to have a good chance of being adopted by a larger community later on. However, NARA should be wary of special-purpose solutions that do not enjoy the benefits of continuous technology improvements.

Sustained improvements in the cost-performance ratio are reflected almost entirely in high-volume products. Examples of high-volume products today are personal computers; ATA/IDE[5] disk drives; compact-disk read-only memory (CD-ROM), CD-recordable (CD-R), digital versatile disk (DVD)-ROM and DVD-R optical storage devices; and Ethernet (10 mega-

[5]The conventional industry shorthand for a family of widely used hard-drive interface standards known as Advanced Technology Attachment (ATA), Integrated Drive Electronics (IDE), Enhanced IDE (EIDE), and Ultra Direct Memory Access (UDMA).

bit [Mb], 100 Mb, and 1 gigabit [Gb]) networks. Each of these products has seen sustained decreases in cost per unit of performance. In contrast, low-volume products rarely exhibit sustainable cost-performance improvement and often disappear rapidly from use. Today's trend in moving to disk from tape for mass storage is an example of what often happens in such cases, with a very high volume product replacing a relatively low volume product.

Despite the anticipated rapid changes in many component technologies, past experience suggests that network interoperability will be relatively stable. Important examples include the backward compatibility among several generations of Ethernet and the persistence of the Internet's Transmission Control Protocol/Internet Protocol (TCP/IP) standards. As networking hardware and software evolve, they are likely to offer backward compatibility of network protocols and interfaces. Thus it will be possible to connect new equipment to old equipment via network hardware of various kinds, and it will be relatively easy to add new capacity to an archival system and to federate archival systems.

Flexible, Modular, and Extensible Design

Effective upgrading of hardware and software to stay on the technology curve of mainstream IT also places a premium on flexibility. At a minimum, the software for the ERA system should be written to be as portable as possible across offerings of different computer and storage vendors. NARA should also seek out products that can be obtained from multiple vendors, so as to avoid being locked in to an arrangement with a single vendor.

Another key to the successful functioning of the ERA over time will be careful modular design, which is critical to allowing components to evolve or to be replaced as needed. It simply takes too long to redesign a monolithic system to exploit technology improvements. Modular design makes a complex problem more tractable by breaking it down into a set of smaller components and enabling the independent evolution of the parts. This type of system is decomposed into separate modules, each with a well-defined interface. Properly designed components can then be replaced with improved versions, while causing minimal disruption to other modules or to the system as a whole. If modules are too large, they become hard to modify or replace. If the interfaces are overly complex or if they permit internal details of a module's implementation to become visible to other modules, the ability to change one module independently of others may be lost.

In addition to being designed for change, the ERA should be designed to be extensible—that is, to be able to adapt to some kinds of changes without requiring modification of the system, at least not immediately. For example, if NARA began accepting a new format, the system should at first accept it as a type "unknown." That way, new records could be accepted without having to wait for the system to be modified. Later, a small, type-specific module could be added to facilitate more sophisticated handling of the type.

LONG-TERM PRESERVATION

The committee's first report, citing the imperative of making significant progress toward implementing an archival system, proposed a pragmatic short-term strategy that included the following: (1) preserving the original bit stream representing a record; (2) converting records to "preferred variant" formats, where available, that are likely to be more readily interpretable in the future; and (3) preserving enough information for records amenable to emulation ap-

proaches so as not to foreclose that option. However, as development of the ERA proceeds, NARA will have to seek out longer-term preservation solutions.

As the committee's first report discussed, a fundamental requirement of the ERA is the ability to store and retrieve the bit streams representing a record and its associated metadata. The requirements for this part of the ERA are fairly well understood (although good implementations will be quite complex, and they will certainly require refinement and evolution over time). In addition, technologies exist or are being developed to satisfy many of these requirements, and research projects are already reaching advanced states.

As a general rule, one keeps the original bits,[6] since this allows for using the original viewer if that is still possible, and it ensures that all information that was in a record is still there (even if interpreting it may be problematic). Also, the bit stream resulting from a reversible transformation applied to the original bits is as good as the original bit stream, provided that the transformation is error-free. The transformation of an uncompressed bitmap image into a run-length encoding produces a file equivalent to the original bit stream, assuming, of course, that the technical metadata identify or define the new representation correctly.

The lowest level of a system architecture for a digital archive is a storage component that sees the bit streams as objects. This component supports a simple object interface used to create, delete (in cases where a specific retention period applies), and retrieve an object. Many of the design issues associated with object storage were addressed in the committee's first report. In practice, good implementations will be quite complex and they will certainly evolve. Therefore, one of the main characteristics of such a subsystem should be its modularity, enabling components to be upgraded without affecting other parts. There is also a growing commercial interest in subsystems with such object stores. This interest is driven by such developments as new regulations that mandate long-term retention.

However, ensuring that bits can be understood in the distant future is a much more difficult problem than are issues involving object storage. Various approaches have been developed,[7] none of which is suitable in all cases. Moreover, one cannot know with any great certainty which proposed methods for preserving the interpretability of digital objects into the distant future will subsequently be deemed successful, much less prove relatively cost-effective.

Preservation of records will thus require selecting the appropriate preservation techniques for particular types of records. An archive can and should, of course, use different approaches for different types of record, and it can and may simultaneously and in parallel pursue several approaches for a particular type of record. Electronic records archivists will need to understand the range of available preservation techniques and make judgments about which are appropriate.

Because each approach has potential drawbacks and associated costs, compromises will be needed, based on levels of service. A compromise may involve preserving a subset of the

[6]The "bit stream" or "data stream" to be preserved does not include media-specific encoding techniques, checksums, and so forth, that take place at the physical storage layer. The marketplace can generally be relied on to shake out any such "under the covers" concerns—for example, no one today worries about whether (or how) a DVD-ROM reversibly encodes the bits.

[7]There is an extensive literature on the long-term interpretation problem. A compilation of some recent work can be found in Gerard Clifton and Michael Day, 2004, "File Formats and Tools," in *DPC/PADI What's New in Digital Preservation*, No. 9, July-December, Digital Preservation Coalition, York, United Kingdom; available online at <http://www.dpconline.org/docs/whatsnew/issue9.html#2.3>, accessed May 1, 2005.

functionalities associated with a particular record. For example, the archivist may stipulate that the printed representation of a document is all that matters. In that case, saving the picture of the page may be sufficient (assuming that there is enough metadata to support efficient access). For another type of document, the archivist may require a higher level of service by asking that the comments hidden in the file also be archived. Then, in order to be equivalent to the original, the new representation will have to contain the comments. A still higher level of service may require that the interactive environment of the word processor itself be reproducible in the future. Box 1.1 provides some additional examples of levels of service.

For each document type (or document type in a particular collection), the archivist needs to determine the appropriate levels of service that should be used. In the example above, preserving the look of the page and the comments is sufficient if the look and comments constitute the essential information of the document. This means that any additional information (for example, which buttons were used to produce it!) is deemed nonessential, or incidental. Any incidental information may be discarded in the preservation process without any loss of value. Deciding what is essential is crucial to good archiving. A report from the National Archives of Australia says it well: "Determining the essence of records . . . is essential to an efficient, effective and accountable preservation program. Focusing on the essence of a record allows us to clearly state our archival requirements for the preservation of that record and to be held accountable against those requirements."[8]

The archivist has three basic approaches among which to select:

1. Keep the information in its original format and rely on running the original viewer on an emulation of the original system, or running a viewer (or at least a program that decodes the original data) on a stable virtual machine.
2. Keep the information in a standard or intermediate format (i.e., converted once at most) and commit to providing a viewer for that format into the far future.
3. Convert the format into a new one when obsolescence is threatening the readability of the record.

Each option is discussed in more detail below.

One General Approach to Preservation: Keep the Information in Its Original Format

As described below, several approaches to preservation rely on keeping the information from a record in its original format.

Relying on the Availability of the Original Viewer

In the long term the original viewer will no longer work, but it is important to realize that the viewer may still exist in the medium-term. When the original representation is kept, that viewer can be used, and the threat of obsolescence is delayed, possibly by several years. The

[8]Helen Heslop, Simon Davis, and Andrew Wilson. 2002. *An Approach to the Preservation of Digital Records.* National Archives of Australia, Canberra, Australia, December. Available online at <http://www.naa.gov.au/recordkeeping/er/digital_preservation/green_paper.pdf>. Accessed May 23, 2005.

BOX 1.1
Examples of Varying Levels of Service for Records Preservation

- *Preserving text—without presentation.* Assuming that a file is an ASCII text file (original or obtained by converting the document once), metadata can be used to specify how to interpret the bits. The metadata must include the definition of every character, either in English or by showing the bitmap of each character in a common font (the bitmap format itself must be explained). But still, the amount of metadata remains very reasonable.
- *Images.* It may be reasonable to assume that a few formats for images may be used as standards (the commonly used options within the Tagged Image File Format (TIFF) standard are an example) and that such formats will be explained in detail in so many places that decoding algorithms will always exist. However, some special formats will always exist as well, and once many formats exist, some will become obsolete more quickly. In such cases, it is preferable to make sure that the format interpretation is specified in the metadata; if the interpretation is too complicated, it may be better to convert the original bit stream into another one that may be less compressed but easier to explain. An alternative is to store a program executable on a virtual machine; this keeps the original format and only converts it to a simple bitmap image (a *logical view*) for presentation on demand. Note that the logical view can be the same whatever the original format is, greatly simplifying the job of the future user.
- *Data structures in XML—no presentation.* Given the widespread and growing use of XML (Extensible Markup Language), an increasing amount of information will be organized and exchanged as XML structures. Nonetheless, metadata describing the semantic meanings of the tags must be captured.
- *Relational databases.* Multiple levels of service can be identified. The most basic level only preserves the data. It comprises the content of the table in a character form and the schema containing the table and column names, comments on their semantics, units, integrity constraints, and so on. Instructions on how to decode the format in which the elements are stored must also be archived (as a precise explanation or as a program written for a virtual machine).

A higher level of service will request that one preserve the querying and reporting capabilities of the initial system. If these capabilities were simply those of standard SQL (Structured Query Language), nothing other than the data would be needed, since a query component could be written in the future. But modern database systems employ user-defined functions. To be able in the future to ask the same questions as those asked today, the definition of these functions must also be preserved. If there is enough incentive to code these functions for a stable virtual machine, a future system would be able to make use of them. However, as soon as the complexity increases, a full reenactment of the SQL system may be needed, which in turn requires an

metadata play an important role: they must specify (very precisely) the program to be used and all of the parameters that will ensure a faithful interpretation. The preservation subsystem may also contain information on various paths for accessing the right application program. For example, the same application program may run on Windows 2000 or Windows XP. Similarly, the same operating system may run on different hardware environments. Therefore, when a component becomes obsolete along one access path, another path may still be avail-

emulation of the original machine. More generally, an application is often used to present the data (presentation is more than just a literal display of the data in the Relational Database Management System [RDBMS]), accept transactions, and so forth, and thus it embodies considerable information about the meaning of the data. The raw data (or data schema) alone does not specify the meaning; either very careful additional documentation must be preserved or there must be some way to execute or emulate the application at least in part (e.g., for presentation).

Another approach is to consider the database as a vehicle in which records are stored so as to provide efficient access and other capabilities. Instead of archiving the vehicle itself, the records can be rebuilt from the database and archived as data structures.

- *General data structure.* Not all records are relational or XML. Many files may have their own format to represent geographical data, engineering data, or statistical data. Not all files can be converted to XML because, in some cases, the volume of the information in them would increase dramatically. Thus, there is a need to specify how to decode these formats as well.
- *Documents with presentation.* When it is important to preserve the look of a printed document, it is always possible to store an image of the individual pages. In any case, the original bit stream can be archived, together with information on how to decode it. But these formats can be so complicated that only a program can do the job. The program can be written for a virtual machine; it would essentially generate on demand a structure that contains both the tagged data elements of the document and their explicit presentation attributes, down to the individual characters' images. Reconstructing a page is then easy; there is no need for any original word processor or viewer. An alternative is to do the conversion once at ingest time and to archive the resulting structure.
- *Spreadsheet.* A spreadsheet is a document with presentation, so the item above applies. A first level of service may be to save the content as an image. But some of the values appearing in cells are specified as formulas that compute them as a function of the values in other cells. In the next level of service, these formulas could be stored as metadata to convey the mathematical relationships between cells. If the archivist decides that a future user should be able to execute these formulas, then the execution of a program becomes necessary. A virtual machine approach or emulation can be used.
- *Dynamic applications.* Extreme cases such as video games or highly interactive applications may require the archiving of programs to reenact the original behavior. If the code exists in a high-level language, compiling it for a stable virtual machine is possible. But more often the best alternative is to emulate the original hardware and software on which the program ran. Since the emulator may have to be time-sensitive, the implementation would be much more involved.

able. A particular use of such metadata may be to trigger an alarm condition when the last path is in danger of being broken by obsolescence.[9]

[9]This functionality has been implemented in a development system. See R.J. Van Diessen, 2002, *Preservation Requirements in a Deposit System*, Koninklijke Bibliotheek, The Hague, The Netherlands. Available online at <http://www.kb.nl/hrd/dd/dd_onderzoek/reports/3-preservation.pdf>. Accessed May 1, 2005.

Relying on Metadata

In some simple cases, it is possible to include a specification of the process itself without introducing a program.[10] For example, the definition of the BMP bitmap image data type can be explained in a few sentences. Unfortunately, however, processes are generally too complicated to be completely defined by a textual explanation.

Relying on Standards

If it were possible to rely only on a small set of well-established standards, it might be reasonable to imagine that the viewer could be rewritten every time the platform changed. Using standards is certainly good practice. However, the variety of formats that exists today is due in part to the fact that different formats have different performance or ease-of-use characteristics. Users and communities may not be satisfied with tools that they do not think are optimal for their applications and their environments.

Emulating the Original System

The approach of emulating the original system consists of keeping the original executable program that was used to create and/or manipulate the information in the first place. That program (with its operating system) works only on the original or an equivalent machine. In the future, the only way to execute the program would be to rely on an emulator.

Initial proposals for emulation suggested that a description of the original machine architecture be archived in all its details so that an emulator might be written when needed. This suggestion is hard to carry out when the expertise on the original machine has evaporated or when the original machine is not available for testing. An alternative to this approach is to write some kind of emulator specification early, when the expertise exists, and to use that information to generate an emulator when the new machine on which it would run is known. A second alternative is to rely on a stable virtual machine,[11] allowing an emulator of the original machine to be written—and debugged—at the time the original machine is known. In the future, a virtual machine emulator will make it possible to execute the emulator of the original machine, providing an original machine that then executes the original application code. The metadata must simply contain a user's guide on how to run the program.

Preserving the execution of the original program is justifiable when reenacting that program's behavior is of interest, but it is overkill for data archiving. In order to archive a collection of pictures when only the pictures themselves are of interest for posterity, it should not be necessary to save the full system that enabled the original user to create, modify, and enhance the pictures. To read an e-mail, it is not necessary to reactivate the whole e-mail system! There is another drawback: in many cases the application program may display the data in a certain way (for example, a graphical representation) without giving explicit access to

[10] J. Rothenberg. 1999. *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation*, Council on Library and Information Resources, Washington, D.C.

[11] R. Lorie. 2001. "Long-Term Preservation of Digital Information." Paper presented at Joint Conference on Digital Libraries, Association for Computing Machinery/Institute of Electrical and Electronics Engineers, Roanoke, Va., June 2001.

the data themselves. In such a case, it is impossible to export the basic data from the old system to the new one, making the reuse of the data all but impossible.

Relying on a Stable Virtual Machine

The method of preservation that relies on a stable virtual machine consists of archiving with the original file, a decoding program *P*, which understands the internal structure and presents the information in a logical view (a hierarchical structure of tagged elements similar to those used in XML). That program *P* is written for the virtual machine. In the future, a restore-application program reads the bit stream and the program *P*; it passes *P* to an emulator of the virtual machine that executes it. During that execution the data are decoded and returned to the client according to the logical view. The schema of the view (the definition of the tags) can also be archived and retrieved by using a similar technique.

Note that in addition to decoding the data structure, the program *P* may perform any arbitrary code on the data. For example, some data elements returned in the logical view can be obtained by computation on other data elements. If the virtual machine provides a minimal communication mechanism, the program *P* can receive parameters from the future environment and compute a new object content dynamically.[12]

A Second General Approach to Preservation: Convert Once to a Standard or Canonical Format

Relying on Metadata

A format can be converted once into a new format. It is quite possible that the new format is much simpler to explain than the original one was. This may be a reason to convert once, probably at record ingest time. On the other hand, the new format may be less efficient (with less compression, for example). There is also a risk of error or loss of features in any conversion. The trade-off must be carefully evaluated. However, converting once does not mean that the original file cannot be kept as well.

Converting the data (once) to XML is starting to be recognized as one approach to the preservation problem.[13] Although it is true that XML can play a very efficient role for purposes of preserving information, it is not by itself a panacea. XML is a convention for identifying data elements in a character string through the use of tags (the particular tags can be chosen to best fit the particular application). XML per se does not specify what to do with the data; in particular, it does not say how the data should be presented. However, the tags provide a way for other programs (such as style sheets), associated with the XML technology, to take care of the presentation by providing a syntax to specify the presentation attributes for each tag (for example: title in Helvetica font, 18-point, centered; author in Helvetica, 12-point, centered; and so on). But XML does not help in preserving the execution of style sheets for the future.

[12]Raymond Lorie and Raymond van Diessen. 2005. "Long-Term Preservation of Complex Processes." Paper presented at Image Science and Technology Archiving Conference, Washington, D.C., April.

[13]For example, the National Archives of Australia has developed an open source software package that provides a plug-in architecture to convert a range of file formats—such as Microsoft Office and OpenOffice formats, relational databases, JPEG, GIF, TIFF, PNG, and BMP—to XML representations for longer-term access.

Relying on Standards

The fact that a format can be converted once into another format is very useful if the other format is so standard and stable that a viewer will remain available "forever." Desirable properties of such standard and stable formats include a public definition, broad applicability, compatibility with prior versions, and a slow rate of change. NARA has played an important role in the development of one such format—the draft PDF/A standard—and can continue to help promote the development and adoption of formats that have attractive archival properties.

Relying on a Stable Virtual Machine

As explained above, a virtual machine can be used to preserve the decoding of an original file. The same applies to a file obtained by an initial conversion of the original file onto a new format. This may be appropriate when tools are available for extracting all essential information from the original document and storing it in a new format that is simpler to decode.

A Third General Approach to Preservation: Migrate the Format into a New One When Obsolescence Threatens the Readability of the Record

The third general approach to preservation is what is generally understood by the term "migration."[14] Each time the configuration of hardware or software is changed in a manner that affects the interpretability of a document file, the data and/or the program must be altered. Conversion poses some special challenges for several reasons: (1) it imposes the ongoing burden of assessing the interpretability of formats contained in the archive; (2) it has unbounded costs, because there is no end to the conversions that need to be done; (3) it may introduce errors at each conversion; and (4) features of a record may be lost as they are mapped from one version to another.

The third consideration—the possibility of introducing errors—is particularly significant. As with other aspects of archive management, human inspection, except on a very sparse sample, will be prohibitively expensive. Even if the conversion failure rate is very low, the consequence of automatic conversion and little or no human checking is to place the burden of discovering errors and recovering from them on the ultimate users of the record. This eventuality is likely to lead the user to fall back on one of the other preservation approaches described above.

Conversion would be more attractive in some particular scenarios—for example, if the original form of a record is in some version of a very popular, well-characterized format such as PDF (which also happens to be relatively stable, in the backward-compatibility sense). When that record is migrated to succeeding versions of PDF and then into successor formats, each of these conversions is likely to be highly accurate because society in general invested a great deal in their being so at the time. Moreover, because all earlier (con)versions would have been kept in such cases, there would be a window of opportunity (while interpreters for the

[14]Note that the term "migration" is sometimes defined also to include the refreshing of bit streams onto new storage media. However, this refreshing is, of course, a fundamental requirement for all three approaches described above.

last format or two are still available) in which errors could be corrected. (And since "live" interpreters would be available for both original and new formats at the time, at least some form of automatic error checking would be possible, such as comparing resulting bitmaps.)

A variation on this strategy is migration on access, which postpones the conversion until a user requests a record. This approach has several possible advantages. For example, deferral can save on resources when most content is rarely accessed; additionally, on-the-fly conversion makes it more likely that the most up-to-date conversion software will be used.[15] However, deferral pushes the problem to a time when data-type obsolescence would make it much more difficult to develop or validate a converter. Also, the advantages of migration on access are reduced in proportion to the fraction of data types that end up being accessed in the future.

GROWING USER EXPECTATIONS

Users of online services, such as e-commerce sites, Web search engines, and the like, have developed expectations for near-immediate access to vast data sets. Users' expectations for the ERA will include the following:

- *Easy, online access to entire collections.* Users will expect the full contents of digital collections archived by the ERA to be searchable and accessible online. They will expect access to be available on an unmediated basis—that is, without going through an archivist—though they may sometimes seek such assistance. People increasingly expect to be able to locate government information online, and the Internet is rapidly becoming the primary channel for accessing government information.[16]
- *Free and anonymous nominal access.* As with other online government information, users will expect to be able to search for and access information archived by the ERA without having to pay fees or to register. Users will be more willing to pay fees for bulk downloads and other such services.
- *Speed*. Users have come to expect near-instantaneous response times even when searching databases containing many billions of items (e.g., Web search engines and e-commerce sites). Users are increasingly intolerant of slow response caused by high demand; those providing Web services today carefully architect and provision their systems to minimize response time. Similar requirements are being placed on government systems.[17] Capacity planning, performance measurement, and production engineering will need to be done on an ongoing basis in order to provide a particular level of quality as public demand scales up. Of course, not all access can be provided on an instantaneous basis: much richer data types and more sophisticated approaches for searching records may require much more processing time.

[15]David S.H. Rosenthal et al. 2005. "Transparent Format Migration of Preserved Web Content." *D-Lib Magazine* 11(1). Available online at <http://www.dlib.org/>. Accessed May 1, 2005.

[16]John B. Horrigan and Lee Rainey. 2002. *Counting on the Internet.* Pew Internet & American Life Project, Washington, D.C., p. 9. Available online at <http://www.pewinternet.org/reports/pdfs/PIP_Expectations.pdf>. Accessed May 1, 2005.

[17]For example, the FirstGov search engine was specified to have a response time under 5 seconds. William Matthews. 2002. "Vendors Vie for New FirstGov Contract," *Federal Computer Week*, January 21. Available online at <http://www.fcw.com/fcw/articles/2002/0121/news-first-01-21-02.asp>. Accessed May 1, 2005.

Nor can prebuilt indexes or other specialized access paths be made available for all types of records.

- *High levels of availability*. Users have come to expect the generally high levels of availability that major commercial services provide by investing substantially in redundant systems and network connections.
- *A well-designed user interface.* Commercial services offer well-designed user interfaces that are continually being refined. Users will expect the same of the ERA's interface.
- *Federated search.* Users are beginning to expect services that allow them to search across multiple sources of content. For example, search engines such as Google provide a federated search across the Web, and Fedstats.gov provides a federated search across data from U.S. federal statistical agencies. Users will, for example, increasingly expect to be able to search for government documents and records regardless of which agency has custody of them.
- *Appropriate interpretation and processing, not just access to the raw data—such as conversion from persistent forms into easily usable forms.* If users were able to access the Nixon tapes online today, for example, they would expect them to be available in Real Player or Windows Media Player format.

User demand for certain types of records can also be quite high, exceeding initial projections. A recent instance was the United Kingdom's release in 2002 of the 1901 census data. According to the U.K.'s National Archives, the Web service was designed to support 1.2 million users per day, but demand on the first day was 1.2 million users per hour. The demand continued to rise in the days that followed, and the Web site had to be taken down so that its capacity could be significantly increased.[18]

These are demanding expectations. In the list above, the last item in particular anticipates that users will expect services that go well beyond the simple delivery of static records on a disk. However, as this report recommends, NARA's priority should be to ingest, manage, and provide basic access to records. Given limited resources, NARA cannot afford to meet all of these expectations for all records—expectations will need to be reconciled against the notion of varying levels of service for different record types. Therefore, third parties should be encouraged to develop user services beyond the basic level provided by NARA, an option discussed in Chapter 3 of this report.

OTHER TECHNOLOGY TRENDS

Full-Content Search and Automated Metadata Extraction

One of the major challenges for the ERA system is that of ingesting the anticipated volumes of records without being bogged down by too much (costly) human effort. Some metadata tell how the records are structured, what kind of records they are, where they came from, when they were created, and so on—information needed both for the management of the records within NARA and to help users find records. Other metadata, such as subject

[18]Additional information is available online at <http://www.nationalarchives.gov.uk/about/operate/meetings/censusadvisory/17jan2002.htm>. Accessed May 1, 2005.

indexing terms, have historically been assigned by humans and are used primarily for finding rather than management.

Several technologies make it possible to avoid the human effort associated with generating the second kind of metadata. One approach to reducing this burden is full-content searching. There has been considerable progress in this area—that an enormous body of publicly accessible Web pages can be searched effectively by search engines such as Google is a major achievement of relevance to NARA—and there is widespread demand for better searching technologies.

Another technique for increasing the degree of automation possible is the use of automatic metadata extraction. This type of technique involves a wide variety of algorithms and approaches, often referred to as information-extraction techniques, for extracting information from text that is either semistructured or natural language. To take a simple example, such techniques enable the sender, recipient, and subject of a memorandum to be extracted automatically even if these items are not explicitly tagged in the document. Information extraction appears in a number of production systems, including the following:

- *CiteSeer*,[19] in which a Web crawler monitors university Web sites, finds things that it identifies as research publications or technical reports, and extracts information such as the title and author names from the file (usually a PDF and PostScript version of the paper), as well as citations to other papers (at the end of the paper). This system, which has been operational since the late 1990s, is very heavily used in the computer science community.
- *WhizBang! Technologies' job search system FlipDog (subsequently sold to Monster.com),* which crawled the Web looking for job postings and extracted data (e.g., location, salary, and title) from them.[20]
- *Google*, which uses information extraction to answer search queries such as "What is Jupiter?" The first link returned will probably be "Web definitions for Jupiter." Clicking on it displays a list of several-sentence definitions that have been extracted from Web documents using a set of rules for finding things that look like definitions.
- *Shop-bots* , which use information extraction of some kind. In the early days most shopbots used handwritten extraction rules, and many probably still do. However, handwritten rules are fragile, because when a site changes, the handwritten wrappers stop working. Thus, there has been much research on automatically learning to extract information from e-commerce pages. The details of which algorithms are used by which sites are generally proprietary.

[19]See C.L. Giles, K. Bollacker, and S. Lawrence, 1998, "CiteSeer: An Automatic Citation Indexing System," in *Proceedings of the 3rd ACM Conference on Digital Libraries* (DL.98), pp. 89-98, Pittsburgh, Pa., June 23-26; S. Lawrence, K. Bollacker, and C.L. Giles, 1999, "Indexing and Retrieval of Scientific Literature," in *Proceedings of the Eighth International Conference on Information and Knowledge Management* (CIKM 99), pp. 139-146, Kansas City, Missouri, November 2-6; Y. Petinot, C.L Giles, V. Bhatnagar, P.B. Teregowda, H. Han, and I. Councill, 2004, "CiteSeer-API: Towards Seamless Resource Location and Interlinking for Digital Libraries," *Proceedings of the 13th Conference on Information and Knowledge Management* (CIKM 2004), pp. 553-561, Association for Computing Machinery, New York.

[20]Related techniques are discussed in U. Nahm and R. Mooney, 2002, "Text Mining with Information Extraction," in *Proceedings of the AAAI 2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases*, AAAI Press, Menlo Park, Calif.

Another technique for the automated supply of metadata is categorization, which discerns groupings of related content.

Although the techniques discussed here are valuable, they do not always yield correct metadata. The errors that occur will tend to have a minimal impact on the precision (the proportion of items retrieved that are relevant) of finding aids, although they obviously will reduce recall (the proportion of relevant items retrieved by a search).

More complex natural language problems, such as parsing and "understanding," are quite difficult and appear likely to remain so. Nevertheless, the information-extraction techniques that have been developed over a period of several decades have reached a level of maturity allowing their use in production systems when guided by human expertise. Although they do not and will not provide the accuracy of a human, they offer the considerable advantages of improved speed and reduced 0cost that is several orders of magnitude lower than that for manual processing.

2

Reengineering Processes to Meet the Electronic Records Challenge

Chapter 1 describes factors driving the shift to electronic records and the growing volume and variety of such records. Electronic records offer considerable new flexibility with respect to duplication (perfect digital copies can be made readily and cheaply), transfer (records can be transferred electronically using networks or physical media), and physical custody (because storage is cheap, it is feasible to maintain multiple copies). Many electronic records exist within information systems, which makes it possible to capture records and associated metadata systematically, and to automate their transfer and processing. The properties of electronic records support the establishment of new processes and relationships with agencies for scheduling and transferring custody of records. The anticipated volume of records that will be submitted for archiving compels the use of automated transfer and ingest.

This chapter first reviews three National Archives and Records Administration (NARA) initiatives that bear on process reengineering. It then points out a number of additional opportunities for NARA.

RECENT INITIATIVES OF THE NATIONAL ARCHIVES AND RECORDS ADMINISTRATION RELATED TO ELECTRONIC RECORDS

Through three major initiatives—the Electronic Records Archives (ERA) program, the Records Management Redesign initiative, and an initiative on records management under the administration's e-government initiative—NARA has taken a number of steps toward building needed technical capabilities for the archiving and preserving of the federal government's electronic records and toward making associated changes in process.

Electronic Records Archives

NARA's Electronic Records Archives program was begun in 1998. Its purpose was to address the growing volume and diversity of electronic records being used and relied on by

government agencies, as well as to address issues of format obsolescence and preservation. A fundamental objective of the ERA program is to confront the formidable challenge of creating a "system that will authentically preserve and provide access to any kind of electronic record, free from dependency on any specific hardware or software."[1] Early steps in the ERA program included NARA's exploration of possible solutions and undertaking of development projects with partners. The partners included the San Diego Supercomputer Center, the University of Maryland, and the Georgia Tech Research Institute.

In August 2003, NARA released a draft request for proposals[2] (RFP) for the ERA system, "seeking input from industry to ensure that the final RFP clearly communicates the purpose of the acquisition."[3] Before preparing this report, the National Research Council's Committee on Digital Archiving and the National Archives and Records Administration prepared two reports providing advice related to the ERA acquisition: its first report recommended a strategy for engineering and acquiring the ERA using an iterative design approach, and a letter report provided specific comments to NARA regarding the draft RFP.[4] Also in August 2003, the General Accounting Office released a report[5] that "found several deficiencies in NARA's plan for the [system's] acquisition," including the lack of a "concept of operations for the system from the users' perspective" and an "incomplete schedule and process to track the costs of the program."[6]

In 2004, NARA released the final version of its RFP and subsequently selected two contractors to develop designs for the ERA.[7] As NARA observes, much "remains to be done before the vision of ERA becomes a reality. The necessary projects and tasks fall in two large categories: [1] designing the structure and [2] building it."[8] NARA hopes to "have a functional subset of the [ERA] system operational in 2007."[9]

Records Management Redesign

Responding to "technological advances, budget cuts, and other events [that] have significantly changed the environment surrounding Federal records management,"[10] the National

---

[1]See <http://www.archives.gov/electronic_records_archives/about_era.html>. Accessed May 1, 2005.

[2]The draft RFP is available online at <http://www.archives.gov/electronic_records_archives/acquisition/draft_rfp.html>. Accessed May 1, 2005.

[3]See the news release, "National Archives Releases Draft Request for Proposal for the Electronic Records Archives," online at <http://www.archives.gov/media_desk/press_releases/nr03-59.html>. Accessed May 1, 2005.

[4]National Research Council (NRC), 2003, *Building an Electronic Records Archive at the National Archives and Records Administration: Recommendations for Initial Development,* Robert F. Sproull and Jon Eisenberg (eds.), and NRC, 2003, "Letter Report on Building an Electronic Records Archive at the National Archives and Records Administration," October 16, both published by The National Academies Press, Washington, D.C.

[5]General Accounting Office (GAO). 2003. *Records Management: National Archives and Records Administration's Acquisition of Major System Faces Risks* (GAO-03-880), available online at <http://www.gao.gov/new.items/d03880.pdf>.

[6]Diane Frank. 2003. "GAO Sees Electronic Archive Problems," *Federal Computer Week*, August 22. Available online at <http://www.fcw.com/fcw/articles/2003/0818/web-nara-08-22-03.asp>. Accessed May 23, 2005.

[7]The final RFP is available online at <http://www.archives.gov/electronic_records_archives/acquisition/rfp.html>.

[8]NARA's "About ERA" document, available online at <http://www.archives.gov/electronic_records_archives/about_era.html>.

[9]NARA's "About ERA" document, available online at <http://www.archives.gov/electronic_records_archives/about_era.html>.

[10]For more information, see <http://www.archives.gov/records_management/initiatives/rm_redesign_project.html>.

Archives and Records Administration began a redesign of federal records management in 2001. This was done through an initiative with the ultimate goals of (1) documenting the current record-keeping and records use environments in the federal government, (2) using subsequent findings to analyze NARA's records-management policies, and (3) redesigning the scheduling, appraisal, and accessioning processes where necessary.

With respect to the first of those goals, NARA issued in December 2001 a report prepared for NARA by SRA International. The report examined the views and perspectives on record keeping and archiving of a sample of federal employees (Box 2.1). The results of NARA's own records systems analyses—which pointed to the importance of "situational factors" within agency records-management operations—were included in the report, which offered "a number of approaches, options, and interventions that may assist NARA and the agencies to improve Federal [records management]."[11]

Building on the work that produced the 2001 report, NARA issued a draft proposal in July 2002[12] for rethinking the regulations, policies, and processes for managing federal records and for developing a set of strategies to support federal records management. A year later, after "working with stakeholders and [testing] proposals," NARA released *Strategic Directions for Federal Records Management*, its "roadmap for redesigning NARA's records-management program and records management in the federal government."[13]

The *Strategic Directions* report begins by asserting NARA's goals with respect to its redesign of federal records management before moving on to more specific strategies and tactics for reaching those goals. The following are among the report's enumerated strategies:

- [NARA] will demonstrate that effective records management adds value to agency business processes. . . .
- [NARA's] approach to records management will be based on the ISO [International Organization for Standardization] Records Management Standard 15489.
- [NARA] will focus on those records that are essential to the government as a whole for accountability, protection of rights, and documentation of the national experience. . . .
- [NARA will] partner with Federal agencies and others to develop, adapt, or adopt products and practices that support good records management.
- [NARA] will provide leadership, in partnership with other key stakeholders, to focus agency attention on electronic records needs and to guide and support solutions to electronic records issues and problems.[14]

The *Strategic Directions* report also details some of the tactics that NARA will use to pro-

[11]National Archives and Records Administration. 2001. *An Overview of Three Projects Relating to the Changing Federal Recordkeeping Environment: Report on Current Recordkeeping Practices Within the Federal Government* (prepared for NARA by SRA International, Inc.). NARA, Washington, D.C. Available online at <http://www.archives.gov/records_management/initiatives/report_on_recordkeeping_practices.html>. Accessed May 1, 2005.

[12]National Archives and Records Administration. 2002. *Proposal for a Redesign of Federal Records Management*. NARA, Washington, D.C. Available online at <http://www.archives.gov/records_management/initiatives/rm_redesign.html>. Accessed May 23, 2005.

[13]National Archives and Records Administration. 2003. *Strategic Directions for Federal Records Management*. National Archives and Records Administration, Washington, D.C. Available online at <http://www.archives.gov/records_management/initiatives/strategic_directions.html>. Accessed May 23, 2005.

[14]National Archives and Records Administration, 2003. *Strategic Directions for Federal Records Management*. National Archives and Records Administration, Washington, D.C. Available online at <http://www.archives.gov/records_management/initiatives/strategic_directions.html>. Accessed May 23, 2005.

BOX 2.1
Major Findings of Study on Current Record-keeping Practices Within the Federal Government

Following are the major findings from the 2001 study prepared for the National Archives and Records Administration by SRA International, Inc., on perceptions in the federal government with respect to record keeping:

- The quality and success of recordkeeping varies considerably across the agencies studied.
- Government employees do not know how to solve the problem of electronic records—whether the electronic information they create constitutes records and, if so, what to do with the records.
- When agencies have a strong business need for good [recordkeeping], such as the threat of litigation or an agency mission that revolves around maintaining "case" files, then [recordkeeping] practices tend to be relatively strong with regard to the records involved.
- Although records officers and other agency records managers are familiar with the life cycle of records, integration between [recordkeeping] and the business processes of many agencies is distinctly lacking.
- For many Federal employees, the concept of a "record" and what should be scheduled and preserved is unclear.
- [Recordkeeping and records management] in general receive low priority, as evidenced by lack of staff or budget resources, absence of up-to-date policies and procedures, lack of training, and lack of accountability.
- While agencies appreciate specific assistance from NARA personnel, they are frustrated because they perceive that NARA is not meeting agencies' broad needs for guidance and [records management] leadership. Primarily, agencies want timely and responsive guidance and leadership from NARA on current [records management] issues.

SOURCE: National Archives and Records Administration. 2001. *An Overview of Three Projects Relating to the Changing Federal Recordkeeping Environment: Report on Current Recordkeeping Practices Within the Federal Government* (prepared for NARA by SRA International, Inc.). NARA, Washington, D.C. Available online at <http://www.archives.gov/records_management/initiatives/report_on_recordkeeping_practices.html>. Accessed May 1, 2005.

mote these strategies and goals. Those tactics include such things as promoting good communications between NARA and stakeholders, modifying guidance and training programs, providing assistance to agencies, providing agency oversight, supporting the development of records-management tools, and reengineering NARA's own business processes.

Recently, NARA has also issued several white papers in support of the *Strategic Directions* report. The white papers focus on—among other things—advocacy, appraisal policy, certification of training, custody, expansion of the records-management training program, inspections and studies of records management, and resource allocation.[15]

[15]The white papers may be found online at <http://www.archives.gov/records_management/initiatives/white_papers.html>. Accessed May 23, 2005.

The Electronic Records Management Initiative

NARA's Electronic Records Management (ERM) initiative was launched as part of the Bush administration's e-government agenda. This initiative is intended to provide "policy guidance to help [federal] agencies . . . better manage their electronic records, so that records information can be . . . used to support timely and effective decision making, enhance service delivery, and ensure accountability." The ERM "will provide the tools that agencies will need to manage their records in electronic form, addressing specific areas of electronic records management where agencies are having major difficulties."[16]

The ERM Initiative was started with the following three goals in mind:

1. To integrate e-records-management concepts and practices with comprehensive information-management policies, processes, and objectives in order to assure the integrity of e-records and information;
2. To employ ERM to support interoperability, timely and effective decision making, and improved services to customers; and
3. To provide the tools for agencies to access e-records for as long as required and to transfer permanent e-records to NARA for preservation and future use by government and citizens.

As the managing agency for the ERM initiative, NARA has organized the project into four issue areas:

1. Correspondence tracking,
2. Enterprise-wide ERM (to provide tools for agencies to use as they plan and implement ERM systems),
3. Electronic information-management standards (to promote government-wide use of the Department of Defense 5015.2 Standard for Records Management for federal agencies), and
4. Transfer of permanent electronic records to NARA (to address an expansion of both the number of formats that NARA can accept and the media and techniques that can be used by federal agencies when transferring their permanently valuable electronic records to NARA).

The overall ERM initiative includes a number of deliverables. For example, NARA has expanded the acceptable methods for transferring electronic records to NARA, and it has issued guidance for transferring permanent electronic records of the following types: e-mail records (including attachments), Web content records, scanned images of textual records, digital photography records, digital geospatial data records, and records in portable document format.

[16]NARA's overview of the Electronic Records Management initiative, available online at <http://www.archives.gov/records_management/initiatives/erm_overview.html>. Accessed May 23, 2005.

PROCESS REENGINEERING TO REALIZE ERA GOALS

Outlined below are broad areas of process reengineering that will be important to the success of the ERA and, more generally, to NARA's success in handling electronic records.

Automated Ingest

Existing processes for ingesting records rely heavily on manual processes. Paper records naturally require a good deal of manual processing to acquire and appraise. Current procedures for transferring electronic records also involve significant manual intervention, starting with the effort associated with creating and organizing metadata into finding aids and the transfer of physical media or the initiation of a network file transfer. Similarly, appraisal and description require significant manual effort, including such steps as manually examining record structures and capturing metadata.

As outlined in Chapter 1, the volume of electronic materials will grow substantially and rapidly. At the same time, a major increase in NARA's budget for archiving appears unlikely. The only possible resolution in order for the ERA system to be able to handle the anticipated volume of records is to make a major shift to automation. Fortunately, it is possible to make this change by improving how records are delivered to NARA, by establishing procedures and systems to ingest new records automatically on a routine basis, and by capturing metadata at record creation and using automatic techniques to help fill in gaps.

To make the automation effort successful, it will be critical to establish how to handle errors that occur in the ingest process. This is important because the costs of exception handling are a common reason for the failure of automation. Careful consideration should be given to the amount of effort that should be expended on trying to fix problems, on the extent to which the responsibility for resolving problems can be pushed back to the originating agency, on ways that errors can be corrected in the future when they are discovered by archives users, and so forth.

Even as NARA develops capabilities and processes to handle new records through automation, a considerable backlog of electronic records will continue to grow before these new measures take full effect. To avoid being swamped by the backlog, triage measures may be needed. These could include avoiding significant manual processing, preserving the obvious artifacts that would support the extraction of automatic metadata in the future (e.g., when accepting e-mail, it is important also to receive the mapping from e-mail address to personal directory information), and using automatic metadata extraction (and accepting its shortcomings).

Early Processing or Ingest of Records

The ingest of electronic records decades after their creation can be problematical—obsolete data and metadata formats may be difficult to interpret, and it may be quite difficult to modify the systems in which the records reside so as to facilitate ingest. When the originating systems are still active, preservation problems can be tackled before obsolescence.

One way to mitigate problems at record ingest is to conduct dry-run testing (running through all of the steps that would be taken at ingest, conducting the dry run over full sets of records, or using a sample as a spot check). Such testing allows a number of potential problems to be uncovered and resolved, including the following: Are the data and metadata for-

mats known, and can they be processed by the ERA's (automated) ingest process? Is appropriate information as to the record's authenticity and integrity available? Are access and transmission procedures in place at the creating agency and at NARA to effect ingest? This type of system debugging serves as an early warning of problems.

Another option is the early capture of records in anticipation of future custody transfer to NARA. Unlike paper records, electronic records offer considerable new flexibility—copies are readily made and cheaply stored. Early capture is best done by designing systems that generate permanent records so that records are created archive-ready at birth—automatically recording metadata, providing users with very easy ways to create metadata manually as records are created, and designing systems that produce outputs that are easy to interpret, self-describing, and otherwise well suited for long-term preservation.

"Pull" technologies—such as automated download of records and Web harvesting—are one way to routinely obtain records earlier in their life cycle. This technique is not limited to publicly accessible records. For example, an agency could provide a restricted-access Web service for NARA to use to extract records as they are created or to search the agency's internal record-keeping system. The resulting archival copies might also provide additional incentives for agencies to support the creation of archive-ready records and their delivery to NARA, since these records would then serve as off-site backups of agency operational systems.

Thus, there are two almost-identical techniques: a passive one in which NARA is sent records early, on a provisional basis, for purposes of a dry run; and an active one in which NARA extracts and ingests records. Both give early warning. The active mode is probably harder for agencies (and NARA) to implement, but it gives NARA the greatest visibility into the records that it will have to handle. The active approach also provides the potential incentive of NARA's copies serving as a backup.

Appraisal, Selection, and Description of Electronic Records

Archival practice was developed primarily to handle paper and other tangible-media records. However, several trends over the past three decades have made it more difficult to carry out traditional records scheduling and disposition activities:

- New ways of working that supplant the traditional organizational hierarchy with greater use of teams, task forces, and so on that are project- or goal-oriented, rather than procedure-oriented;
- The reduction in administrative and clerical staff responsible for organizing, filing, and retrieving records and developing retention and disposition schedules, and the resulting transfer of responsibility for record management to end-user agency staff;
- The preference for electronic communication and record keeping over paper-based systems; and
- The dearth of useful tools that can integrate records management into end-user tasks and activities.

At the same time, several technology trends compel a rethinking of selection, appraisal, and description in two major areas:

• *First, the granularity of appraisal should be larger*. The appraisal of physical records was driven in part by a need to reduce significant storage costs in an archive. But with ever-declining storage costs, the opposite is the case with digital information. Investing time in reducing a set of records by 10 percent or 40 percent or even 60 percent costs more than it saves, because these decisions require human intervention and professional judgment and research. Also, records are more interconnected when they are digital. The information or evidential value of a set of records considered in isolation may not be obvious, but one set of records may be crucial for understanding or using another set.

Rather than examining sets of records carefully and marking only a select few for retention, it may become more expedient and more appropriate to retain larger "chunks" of records, The committee is not advocating a "save everything" approach, but a rethinking of selection criteria. For example, it may be better to preserve entire classes of electronic records (e.g., all of the records from certain offices, all of the records covering function *X* in both the federal and regional offices, and so on) that would not have been feasible for paper records, even though some records with little permanent value may be intermingled. NARA appears to be moving in this direction in the Records Management Redesign initiative, with its emphasis on functional appraisal, which implicitly is more broadly granular.

Consider a specific example. For the material that will be included in the U.S. Department of State's new SMART system, NARA currently has more than 2,000 retention and disposition schedules for series of records with various retention periods. It might be much more cost-effective simply to keep everything from this system, or to replace the current fine level of appraisal and disposition with one that has fewer than a dozen schedules.

• *Second, archives should place less emphasis on manual record description of records and on creation of finding aids and more on automated tools for improving access.* The characteristics of electronic records permit new access methods that can decrease the burden on archivists to describe records and create finding aids. Agencies can supply records in a structured format and attach metadata. Full-content search on the records and associated metadata can supplement—or even supplant—manually generated finding aids. Finally, agencies themselves create search and access tools for some of their records. Such tools might be maintained by the originating agency (perhaps as part of the sort of agreement to distribute archiving responsibility, as discussed below) or adapted for use by NARA.

Another way to think about the trade-offs discussed here is that it may be advisable in many cases to make access for end users (when and if it occurs) potentially more difficult (by increasing the total volume of records to be examined and describing records in less detail) in exchange for making ingest cheap (by reducing the burden on the archivist to select and describe materials). Of course, as tools such as full-content search improve, the burden on the end user will be reduced.

Delivering Records Archive-Ready

Electronic records transfer to NARA and their ingest are greatly eased if records can be supplied to NARA "archive-ready." Archive-ready records, at a minimum, are packaged with the metadata required for them to be transferred and otherwise processed using largely automated processes. Preferably, archive-ready records would also be provided in structured, standardized formats approved by NARA as amenable to long-term preservation.

The closer to record creation that archive readiness is established, the more likely it is that

the appropriate metadata will be captured. This means that less work (and cost) will be incurred in adding the information later. Ideally, agencies would design systems so that records are created archive-ready at birth. This condition would occur through automatic recording of metadata, agencies providing their own staff members with very easy ways to manually create metadata as records are created, and so forth. Alternatively, agencies can add necessary metadata at some point after creation, but then still deliver records to NARA archive-ready.

If a large fraction of the records delivered to NARA are to be archive-ready in the long run, archiving considerations will have to become part of the government software procurement and development process. This is especially true for those systems that produce or are deemed likely to produce permanent records. Successful implementation of the ERA system may also require NARA to become more actively involved in the establishment of standards used in constructing federal systems, such as those governing the formats used to represent records and associated metadata, and for NARA—or other bodies responsible for information policy, such as the Office of Management and Budget—to establish guidance to agencies and auditing and enforcement measures as necessary.

As a practical matter, these efforts should link archive-ready concerns with agency interests in supporting the agencies' own record-keeping operations. Experience shows the difficulty of implementing systems that depend on user compliance with externally imposed requirements that necessitate effort and do not support the user's internal business needs. As a result, requirements should support agency record-keeping needs, and implementing the records should be realistically possible. Experience also shows that schemes that depend on user-supplied metadata are unlikely to succeed, meaning that software and systems that minimize additional manual contributions of metadata are more likely to succeed.

Close coordination will also be needed between the Records Management Design initiative and the ERA program, because if changes along the lines of those envisioned in the RMI do not occur, the flow of records into the ERA will be seriously impeded. The amount of work that NARA would then need to do to make the records useful would be unsustainable.

Metadata Collection

As discussed above, metadata describing electronic records can and should be created with as little recourse to manual description as possible. It is useful to distinguish three ways in which metadata can be collected automatically:

1. *Metadata can be collected as a matter of course when a record is created (or by virtue of the series or collection in which it lies).* For example, a word processor may automatically attach metadata identifying the author and last modification date (assuming that the system is correctly configured with such information).
2. *Metadata can be automatically extracted from a record without any recourse to information extraction or otherwise "understanding" the record.* For example, header information in an e-mail message is readily extracted by parsing the message.
3. *Metadata can be automatically extracted from a record using information-extraction or categorization techniques.* As discussed in Chapter 1, these techniques are imperfect, but they can be expected to improve with time.

The first of these options is relatively easy and inexpensive. It is the preferred approach for authored records. Ideally, agency record-keeping systems would have facilities for automatic capture built in. The second option is also easy and inexpensive for records that have the necessary structure.

The third option should be viewed as a fallback. Even when it is necessary to use automatic metadata-extraction techniques, this is better done early on, by the creating agency, which may be able to use content-specific techniques.

In cases in which NARA will have to resort to automatic metadata extraction using natural language techniques, it will have to accept significant imperfection in some of its metadata—a shift from the traditional mind-set of archival institutions. It will also be necessary for systems and processes to distinguish among different degrees of quality in the metadata, which have varying reliability. Finally, as metadata extraction techniques improve, the metadata can be upgraded (possibly on-the-fly as records are accessed). The ERA should accommodate making such updates to record descriptors.

Cooperative Arrangements with Agencies

The increased flexibility afforded by electronic records opens up a number of options for custody and access in partnership between NARA and federal agencies. The model of having certain agencies retain responsibility for certain types of records is well established; for example, several agencies have been given the responsibility for preserving large scientific data sets. NARA policy has established criteria for determining whether to establish such an affiliated archive.[17] In 2003, NARA entered into a new partnership with the U.S. Government Printing Office (GPO) which makes the GPO an "archival affiliate."[18] The agreement ensures the permanent availability of the documents available today (and in the future) through the GPO Access Web site, which provides online public access to more than 250,000 publications, including the *Congressional Record*, *Federal Register*, and C*ode of Federal Regulations*. The agreement gives NARA legal custody of the documents while delegating to GPO the responsibility for public access and preservation. Users are directed from NARA's catalogs to documents held for access and preservation by GPO.

This approach has several important advantages:

- By leveraging the skills and resources of other agencies for preservation and access, NARA can spend its scarce resources on problems that agencies cannot address on their own.
- Giving an agency responsibility for the preservation and access of records means that the specialized expertise required to understand or manage a complex system need not be transferred to NARA. For example, the U.S. Census Bureau currently provides public access to survey data through a Web interface and supporting analysis software. It would be impractical to expect NARA to develop domain expertise sufficient to operate online access to all forms of government data.

[17]National Archives and Records Administration (NARA). 2003. *Strategic Directions: Custody of Federal Records of Archival Value*. Section 10. NARA, Washington, D.C. September. Available online at <http://www.archives.gov/records_management/initiatives/custody.html>. Accessed May 23, 2005.

[18]See the NARA press release at <http://www.archives.gov/media_desk/press_releases/nr03-60.html> or the GPO press release, "The GPO and National Archives Unite in Support of Permanent Online Public Access," at <http://www.gpo.gov/public-affairs/news/03news46.html>. Accessed May 23, 2005.

- It provides a clear path for NARA to identify permanently valuable essential evidence from a government-wide perspective, and it also enables agencies to make independent judgments as to whatever else they want to retain for their own purposes.
- Diversity of responsibility is created with multiple agencies taking on the responsibility of preservation, and this diversity provides some additional measure of safety compared to that afforded by a monolithic repository.
- By banding together, NARA and other organizations stand a much better chance of spreading widely the standards and tools for preparing records that are archive-ready, of developing affordable software to meet the needs of researchers and other clients of these repositories, and of otherwise making electronic records archives practical in the long run.

Despite these advantages to cooperative arrangements, there is a potential downside to a distributed set of archives. If systems proliferate and diverge, users of affiliated archives will be frustrated by different user interfaces, different access or authorization requirements, and different implementation of services (such as downloading electronic records). To increase the utility to users, the affiliates should federate search and authorization techniques and coordinate their services.

If such arrangements become more common, NARA's role would shift to that of a standards-setter, federator, troubleshooter, and repository of last resort.

3

Partnering with Other Institutions

The National Archives and Records Administration (NARA) need not go it alone in providing services to its users. NARA's work with electronic records affords a number of opportunities to leverage and partner with other organizations, commercial firms, and researchers to provide access to NARA collections. Designing the Electronic Records Archives (ERA) to feature public interfaces can enable much broader use of NARA holdings. The inclusion of appropriate application programming interfaces (APIs) in the ERA design would enable third parties to access records and to provide access services that are layered on top of the NARA-provided records and access understructure. Federation would allow collections held by NARA and other organizations to be accessed as if they were a single collection. This chapter briefly describes ways of partnering with other institutions and technologies that would make such partnerships possible.

APPLICATION PROGRAMMING INTERFACES

An application programming interface allows access to selected system internal data or operations, enabling people to build new interfaces or applications on top of NARA's systems. APIs would allow third parties to do such things as run queries against NARA holdings (catalogs or content) and retrieve individual records. A good example of the potential of open API access is illustrated in the commercial arena by the search service Google. Simple API-level access to its search engine is provided to anyone who registers, letting others spend their own research resources developing applications that might eventually be interesting or useful to Google.

Capabilities that might be built on top of the ERA include the following:

- *Data annotation to add and/or correct record metadata.* Third parties with an interest in particular records could provide services to annotate records—using systems separate from the ERA but accessible to users as if they were part of it. These annotations could be held by

external parties (provided there was a way for third parties to point to ERA records unambiguously, e.g., via a unique identifier), or the ERA could hold and provide access to the overlays.[1]

- *The continued introduction of specialized enhanced capabilities by partners and vendors to support particular needs*. For example, a third party or NARA might supply an interface that provides translations of popular audio recordings (e.g., the Nixon White House tapes) from the format in which they are archived into commonly used formats (e.g., Windows Media or Real Audio). Some of these capabilities might prove useful enough to be folded back into the ERA system proper.
- *Entrepreneurial innovation to provide value-added services, through either APIs or bulk download.* Today, for example, third parties provide value-added access to information in the Securities and Exchange Commission's Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system. Similarly, commercial third parties currently provide enhanced access to old data of the U.S. Census Bureau.
- *Educational materials*. Overlays onto NARA holdings could be a relatively low cost method of dramatically providing a new set of services that satisfy NARA's educational mission. Educational publishers, as well as historians and social scientists, could develop online materials that draw on NARA holdings.

Such capabilities would allow NARA to tap at low cost the resources of third parties in order to provide enhanced services, help fulfill NARA's mission, increase the perceived value of NARA's holdings, and explore new technologies and approaches.

In allowing such capabilities to be built on top of the ERA, NARA will have to consider the problem of how a typical user distinguishes between results that are provided directly from NARA and those produced via a third-party extension. It would probably not be practical for NARA to certify third-party software or services. However, NARA may want to find ways of letting users know when they are using NARA versus third-party software and services, because the guarantees that NARA would make about the authenticity of what is viewed in the two cases would be very different.

Additionally, support for bulk download of records should be provided by NARA to allow third parties to retrieve sets of records against which to run their own full-content searches, perform data mining, and so forth. And an additional design issue that would facilitate third-party access is the adoption within the ERA of a persistent (i.e., stable and durable) record-identifier scheme.

FEDERATION

Federation is a technique whereby a software layer is used to make a collection of relatively different systems—which are often controlled administratively by different organiza-

[1]The overlay concept also has application internally within NARA. For example, Thomson West undertakes to rebuild the full index for the Westlaw databases only every 10 to 15 years. Between rebuilds, it augments the index with new information, some generated automatically, some manually. The system is structured so that a query collects information from all of these sources—the original index and the subsequent overlays. NARA could accept overlaid annotations from partners to add value to its holdings.

tions and often integrated loosely—appear to be a single system.[2] Federation has technical (e.g., the definition of standards for the federation layer and connection to the systems below the federation layer) and organizational (e.g., agreement on standards and the forging of agreements to federate holdings) dimensions.

Applications of federation to the ERA would include the following:

- *Federated access.* There are several reasons why federated access to electronic records collections held by NARA and other organizations would be an attractive capability. First, it would provide a common interface to federal records regardless of which agency happened to have responsibility for managing them, allowing responsibility for electronic records preservation to be distributed among multiple agencies without giving up the advantages of a common access point for users. For example, users could retrieve in a single query records held by NARA as well as agency records that had not yet been transferred to NARA. Federated access would also be valuable, for example, in allowing access to span NARA and presidential library resources. Another example would be enabling seamless access across multiple systems containing archived and active data. For example, someone searching for data about the Snake River would otherwise have to look across many U.S. federal agencies (e.g., the Bureau of Indian Affairs, the Environmental Protection Agency, the Department of the Interior, and the Library of Congress), and the archives and agency records of four states (Wyoming, Idaho, Oregon, and Washington) as well as NARA, all of which might have relevant records. With agreement on how to federate, access could be enabled across NARA and the archives and other holdings of international, state, and local archives, libraries, and other institutions.
- *Federation of archival systems.* More generally, it may prove useful to federate the ingest of records, their access, and other functions across multiple archival systems. This arrangement would provide a way to combine multiple instances of the ERA, allow simultaneous operation of multiple versions of hardware and software, and generally avoid having to maintain a single monolithic archival system.
- *Federated storage.* Placing a federation layer between an archival system and its file storage is a very useful technique. Among its benefits are that file systems at multiple sites can be treated the same by software using the file system, and new implementations of file systems (or storage types) can be added to the system as it evolves. Old file systems, no longer used, can easily be removed.

Federation is a mainstream design approach employed in a number of commercial applications today. Databases are routinely federated, usually at the Structured Query Language (SQL) level, in industry. Federated file systems (the San Diego Supercomputer Center Storage Request Broker[3] is one example) similarly appear in a variety of production systems today. The Z39.50 Information Retrieval Protocol[4] is used to support federated cross-library searches. Following are some other examples of applications that use federation:

[2]By contrast, the term "distributed" usually means multiple instances of identical or similar systems, often under a single administrative control, often tightly integrated. The terms "federated" and "distributed" label end points in a spectrum.

[3]A detailed description is available online at <http://www.sdsc.edu/srb/>. Accessed May 1, 2005.

[4]ANSI/NISO Z39.50—2003 Information Retrieval: Application Service Definition and Protocol Specification.

- *Some comparison shopping systems*. For example, Orbitz federates several airline reservation systems. It is not just a portal providing access to multiple sites, but an actual federation of several airline reservation systems.
- *Identity servers managing identity in computer systems.* In such an application, a (corporate) user logs on to a single identity server; that server then prepares credentials for various information technology (IT) systems when asked by the IT system, using whatever credential types the particular IT system requires. The Liberty Alliance is an industry group working to build federated identity capabilities.
- *New designs federating mail servers.* With such a federation, when one queries a set of e-mail folders (e.g., in a large e-mail archive), the folders residing on different e-mail servers will be searched within their respective servers, but the results will be consolidated. This form of federation involves only a single product, so it is simpler than a federation in which the federated systems run different software (in such a case, the federation interfaces must be carefully defined).

Despite the use of federation in various commercial applications today, there is no common way to federate access to electronic records archives or digital libraries. There is already a significant technology base to draw on, including metasearch technologies and standards such as the Z39.50 Information Retrieval Protocol and the Open Archives Initiative Protocol for Metadata Harvesting.[5] There is, however, no established base of practice for federating access to archives. As NARA and other organizatons start building such archives, federation is a technical area in which NARA might become a leader or engage the research community by informing it of its needs. If NARA is to take advantage of these opportunities, it will also have to address policy issues involved with allowing its resources to participate in federations (that might be run by groups other than NARA).

[5]Additional information is available online at <http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm>. Accessed May 23, 2005.

4

Broadening Research Interactions

The Electronic Records Archives (ERA) program of the National Archives and Records Administration (NARA), currently being undertaken to build significant new capabilities for preserving digital records, has contracted out for two system designs in preparation for the acquisition of an operational system. As a prelude and ongoing complement to procurement activities, the ERA program has sponsored several research and development (R&D) activities (Box 4.1). This chapter examines the role that future research could play in helping NARA better understand and address emerging and future technology issues and medium- and long-term technical challenges.

RATIONALE FOR SUPPORTING RESEARCH

Information technology (IT) research has traditionally been concentrated in several federal agencies, many of which have research as part of their central mission.[1] A recent National Research Council (NRC) report, *Information Technology Research, Innovation, and E-Government*,[2] broadly considers the rationale for IT research by a broader set of federal mission agencies. It concludes that, where possible, government should follow the private sector in designing and

[1]Twelve agencies or major units of agencies were participants in 2004 in the National Coordination Office for Information Technology Research and Development, which coordinates activities and prepares an annual summary of the federal IT R&D budget. These were the National Institute of Standards and Technology, National Oceanic and Atmospheric Administration, Defense Advanced Research Projects Agency, Defense Information Systems Agency, National Security Agency, National Nuclear Security Administration, Department of Energy's Office of Science, Environmental Protection Agency, Agency for Healthcare Research and Quality, National Institutes of Health, National Aeronautics and Space Administration, and National Science Foundation.

[2]Computer Science and Telecommunications Board, National Research Council. 2002. *Information Technology Research, Innovation, and E-Government*. National Academy Press, Washington, D.C.

BOX 4.1
List of Electronic Records Archives Research Projects

The Web site of the Electronic Records Archives program lists the following as current research projects:

- *Persistent Object Preservation.* Work by the National Partnership for Advanced Computing Infrastructure, led by the San Diego Supercomputer Center (SDSC), examining digital preservation methods.
- *Presidential Electronic Records Pilot Operations System*. Work sponsored by the National Archives and Records Administration (NARA) at the Georgia Institute of Technology Research Institute, on archival processing of presidential electronic records.
- *Distributed Object Computation Testbed*. This project, led by the SDSC, brings together supercomputers, large electronic archival technology, and high-speed wide-area networks. Its purpose is to develop solutions for handling massive amounts of data that are stored at sites nationwide and to provide an infrastructure on which to test advanced technologies for the preservation of electronic records.
- *U.S. InterPARES (International Research on Permanent Authentic Records in Electronic Systems) Project*. NARA is a participant in this project, which is studying the problem of authentic electronic records.
- *Grid Technology.* Grid technology is being developed in a joint partnership with the SDSC and the University of Maryland Institute for Advanced Computer Studies. Its purpose is to provide a distributed interoperable network that will provide efficient and effective access to users, regardless of location, and be linked using a logical name space to create global, persistent identifiers.

SOURCE: National Archives and Records Administration. 2005. Electronic Records Archives: Research. Available online at <http://www.archives.gov/electronic_records_archives/research/research.html>. Accessed April 8, 2005.

implementing IT-based services. However, the report also concludes that there are some areas, such as the long-term preservation of government records, in which government requirements differ, at least somewhat, from those in the commercial world. In these areas, the report says, federal agencies should recognize and act on their role as a demand leader.

The NRC report goes on to characterize some of the benefits of research activities for both government agency sponsors and researchers. Government stands to benefit from collaboration between government agencies and the IT research community despite the fact that these two groups are at opposite ends of an extensive supply chain that also includes vendors and system integrators. Although government agencies and researchers may appear to be unlikely allies, they have a shared interest in innovation and in meeting future needs.

Through research, agencies can gain an understanding of emerging and future technologies, and of the technology risks they face. The most effective form of risk reduction occurs when an organization does in-house testing of research prototypes; with in-house testing the organization gains firsthand experience about what is easy and what is not, where the system is reliable and where it fails, and how to run the system on a daily basis. A somewhat-less-

effective form of risk reduction occurs when an external organization develops and tests the prototype; in this case the organization learns something, but most of the expertise about how the system *really* works resides within the external research organization. Agencies also gain an opportunity to influence the trajectory of technologies through engagement with researchers. Research should, however, not be seen as a way of short-circuiting the IT supply chain of researchers, technology vendors, and system integrators, and agencies should not necessarily be early adopters. However, when agencies and researchers work together, the overall risk of innovation in acquisition can be reduced.

Absorptive capacity for research results requires in-house people who can be engaged at a technical level with relevant research communities and who will take results from the researchers, test them in-house, and explain and propagate results within the agency. Mere funding is not enough. The real goal is substantive technical coupling between the research community and the agency that allows the organization to learn about new technologies, to learn more broadly about a research area than it would otherwise, and to understand how to apply research that it is not itself supporting.

RESEARCH MANAGEMENT FOR AGENCIES WITHOUT A TRACK RECORD IN SPONSORING RESEARCH IN INFORMATION TECHNOLOGY

For an agency such as NARA that has neither a long track record of sponsoring IT research nor a large research budget, how can limited research resources be spent most effectively? *Information Technology Research, Innovation, and E-Government* observes that most federal agencies have limited experience in and capability for managing IT research programs. The report goes on to recommend that consideration be given to cross-agency collaboration with agencies that already have IT research programs and/or that share common research interests.

Partners with Research Management Expertise

Several federal agencies, most notably the National Science Foundation (NSF) and the Defense Advanced Research Projects Agency (DARPA), but also the National Institute of Standards and Technology, the National Aeronautics and Space Administration (NASA), the Department of Energy, the National Institutes of Health and its National Library of Medicine, the armed services laboratories, and others, have extensive expertise in the management of IT research projects. The selection of specific research proposals and routine management of the research process—for example, selecting among peer-reviewed IT research proposals or maintaining an arm's-length relationship with research organizations—are capabilities that these agencies have developed over time as they have gained experience, identified and worked with particular research communities, brought in researchers to manage programs, and otherwise developed research management talent. These organizations maintain relationships with varied research and user communities, which gives them easy access to a wide range of expertise and peer review when needed.

The agencies referred to above also have experience in launching research programs in new areas, including the building of a research community around a set of technical topics that were not previously identified as related or important. This process is more complicated than simply stating a problem and looking for people to work on it. Researchers will not be drawn

to work on problems that are not well thought out, for example, or that are shallow or short-term enough that they do not lead to significant research results (e.g., sufficient for the promotion of academics or the funding of start-up companies). Strong, successful research communities often are developed by several customers and funders. The digital libraries program managed by NSF is a highly relevant example of successful program development and community building.

Exchanging staff for limited assignments is a way of both learning about relevant research and letting others learn about one's own research needs. NARA staff members could spend profitable 6-month assignments in other research organizations to learn what is on the cutting edge and let the researchers know what NARA's problems are. Similarly, members of research organizations could spend profitable 6-month assignments inside NARA learning what NARA is really up against and offering suggestions on how to apply what they know to these problems.

Partners with Shared Research Interests

Many of NARA's most basic needs for digital preservation and access are shared to varying degrees with organizations such as the Library of Congress, the National Library of Medicine, NASA, the National Oceanic and Atmospheric Administration (NOAA), and the intelligence agencies, all of which must maintain indefinite access to digital records. The Library of Congress has established the National Digital Information Infrastructure and Preservation Program (NDIIPP) to work collaboratively with federal agencies, research libraries, and other organizations that have an interest and expertise in digital preservation. It has partnered with the National Science Foundation to establish a research grant program focused on digital preservation and has launched joint research activities with four universities to investigate various approaches to digital preservation using a test collection.

A number of the technical problems faced by NARA are likely to be faced by other agencies as well. For example, semiautomatic redaction or declassification is a topic that could be investigated with partners such as DARPA or the intelligence community's Advanced Research and Development Activity (ARDA). Another organization with shared interests in preservation is the Department of Veterans Affairs, which is currently developing the HealtheVet-VistA program to provide long-term access to veterans' electronic health records (Box 4.2).

Partnership on Research-Oriented Prototypes

As NARA becomes engaged with a broader range of research communities, it will be natural for it to participate in research-oriented prototypes. The National Science Foundation's Digital Government program has, for example, fostered close relationships between universities and agencies such as the Bureau of Labor Statistics, the Census Bureau, and the Environmental Protection Agency. NSF's Digital Libraries program has fostered similar government and academic partnerships. These partnerships give the research community access to real data and users and thus an opportunity to receive real-world feedback about their research. In return, government agencies gain experience with new technologies before the innovations become commercially available, and the agencies develop long-term relationships with outside experts. Such partnerships and in-house testing of research prototypes should eventually become an important part of NARA's research portfolio.

BOX 4.2
HealtheVet-VistA

HealtheVet-VistA is a program currently under design. It will create an online environment in which veterans' electronic health records, medical images such as x-rays, pathology slides, scanned documents, results of cardiology examinations, wound photos, and endoscopies, among other records, are stored for ready access.

Currently, the Department of Veterans Affairs (VA) treats more than 3 million patients annually in more than 1,300 health units, 163 hospitals, and 850 clinics. The Committee on Veterans' Affairs of the U.S. House of Representatives has mandated that the VA retain all veterans' health records (which are in electronic form) for 75 years after the death of veterans. Given technology obsolescence, this mandated retention period creates an enormous challenge for the VA. The VA also faces a comparable challenge in standardizing veterans' electronic health records across all of its components and in implementing "upstream" creation and capture of veterans' electronic health records that facilitate access for 75 years or more.

Setting aside the commitment of the National Archives and Records Administration (NARA) to permanent storage (i.e., forever) of selected electronic records created by federal agencies, the VA and NARA face a comparable challenge over the next hundred years in ensuring long-term access to electronic records. The two agencies also share common interests in approaches, tools, and procedures to support long-term access to electronic records.

ENGAGING THE RESEARCH COMMUNITY THROUGH MEANS OTHER THAN FUNDING RESEARCH

NARA can engage relevant research communities in a number of ways other than direct sponsorship of research. They include the following:

- *Hosting workshops and seminars.* Workshops and seminars bring researchers and NARA employees together, allowing a two-way exchange of ideas and problems. The ERA program took advantage of this opportunity by holding a November 2004 conference, called "Partnerships in Innovation: Serving a Networked Nation," organized in conjunction with the University of Maryland's Institute for Advanced Computer Studies.
- *Hiring students.* Another way of engaging research communities is to seek out and hire their students. This is well understood in industry as a very effective approach to technology transfer.
- *Suggesting important problems and providing access to artifacts (data, software, and system capabilities) and subject-matter experts.* As the NRC report *Information Technology Research, Innovation, and E-Government* observes, working on government IT problems offers researchers access to applications with a "richness and texture often lacking in the laboratory" as well as other potential benefits.[3]

[3]Computer Science and Telecommunications Board, National Research Council. 2002. *Information Technology Research, Innovation, and E-Government.* National Academy Press, Washington, D.C., pp. 12-13.

The opportunity of providing researchers with access to artifacts is discussed in more detail in the following subsection.

Providing Access to Artifacts

As a complement to sponsoring research, NARA can also engage the research community by providing data. This opportunity represents a very low cost yet effective way of engaging substantively with the research community. Academic researchers are often starved for data; they have good ideas, but no way to test them against reality. Providing interesting data in a form that the research community can use can be enough to get the best people in the world working on a problem, essentially for free. For example, Reuters created two databases of categorized newswire articles for research use and made them easy to acquire. Almost every text classification algorithm developed in the preceding decade was tested on these data sets, and hundreds of research papers have been published about them, none paid for by Reuters. Similar stories can be told about data sets from news organizations such as the Wall Street Journal, AP Newswire, Ziff-Davis, and the Financial Times, and about medical data sets based on the National Library of Medicine's MedLine system.

There are many vehicles for providing data to the research community. Since 1992 the National Institute of Standards and Technology (NIST) has created a variety of widely used data sets that have focused research attention on a diverse set of problems, often of interest to the U.S. intelligence community. Some universities, for example the University of California at Irvine and the University of Pennsylvania, have specialized in creating and becoming long-term distributors of research data sets. Partnering with an organization that has prior experience in this area, such as NIST, the University of California at Irvine, or the Linguistic Data Consortium (LDC) at the University of Pennsylvania, would help NARA learn how to create data sets that are engaging and have long-term value to various research communities.

Chapter 3 in this report points out the benefits of providing an application programming interface (API) that gives direct access to some of the ERA's digital archive so that third parties can develop new tools. This approach is particularly applicable to research initiatives. Google, for example, allows a simple API-level access to its search engine to anyone who registers. The advantage to Google is that others can spend their own research dollars to develop applications that might eventually be interesting or useful to Google. API-level access can also provide access to users, enabling researchers to conduct experiments with live populations in real time—for example, by varying system characteristics and observing user behavior. Such access would be extremely attractive to a variety of computer science and social science research communities, and it could be provided by NARA at relatively low cost relative to the benefits obtained.

RESEARCH CHALLENGES FACING THE NATIONAL ARCHIVES AND RECORDS ADMINISTRATION

The research problems facing NARA—and appropriate research strategies—fall into two categories:

1. *Problems shared with other organizations.* Many of NARA's technical problems are the same as those faced by designers and operators of any large digital library or repository. For

dealing with these problems, NARA's greatest leverage will come from drawing on research sponsored by other organizations, learning about best practices, inducing others to work on problems of interest to NARA by offering corpora of content that researchers can use to test new ideas, and participating in joint research programs with other federal agencies and organizations that face similar challenges. Wherever technologies may be shared with other applications, partnerships to jointly address shared problems will help stretch limited research resources.

2. *Problems specific to government archives and similar institutions.* A few problems are specific to the preservation of records or otherwise unique to NARA and may need separate research thrusts. Working on these more specialized problems may require specific engagement with existing research communities. Partnership with agencies that have greater experience in managing IT research programs is also likely to be the most effective mechanism for addressing this class of research problems.

What are some of the specific research problems that NARA faces? The organization must develop its own list, but it need not start from scratch in doing so.[4] The subsections below provide some illustrative examples of research problems in each class.

Research Problems Shared with Other Organizations

Following is a list of examples of research problems that NARA shares with other organizations, together with appropriate strategies.

- *Automatic classification.* Traditional archiving processes rely heavily on controlled-vocabulary metadata, which is usually assigned manually. Large-scale use of automatic categorization to assign metadata provides a way to handle the expected flood of digital records in a cost-effective manner. There has been significant progress in this technology area during the past decade, and there is now a greater understanding of what makes problems "difficult" and "easy." Automatic categorization, which is used routinely in a variety of commercial settings, is often as accurate as human categorization is. Automatic categorization must become a core competency for NARA. Human effort should be applied only when automatic solutions are inadequate. The use of this technology can be broadened to related problems such as filtering records for retention and distinguishing between federal and presidential records.

Engagement with leaders in the research community would provide NARA an opportunity to better understand the state of the art. NARA's contributions in this case might well not involve direct support for research; NARA could make significant contributions to the research community by making corpora available, stating its own requirements clearly, and learning from research sponsored by other agencies.

- *Scholars' tools for searching large collections of records.* Some shortcomings of record classification can be remedied by increasingly powerful search methods. While a casual searcher

[4]See, for example, Library of Congress and National Science Foundation, 2003, *It's About Time: Research Challenges in Digital Archiving and Long-Term Preservation,* Margaret Hedstrom (ed.), Library of Congress, Washington, D.C., p. ix, available online at <http://www.digitalpreservation.gov/repor/NSF_LC_Final_Report.pdf>; accessed May 1, 2005.

may be content with a result that is "good enough," scholars want better tools and are likely to be continually pressing the limits of technology. For example, Google already offers a "Google Scholar" search designed to find online scholarly material, and libraries and technical literature publishers are beginning to offer customized searching. Many of these emerging techniques will probably be useful to scholars searching the ERA, but the archives may pose some unique problems. This is a case in which NARA could stimulate a research community with a database of archival records and a set of demanding users with specific needs.

- *Preserving computer programs.* Improved capabilities for preserving software may be of general interest to the community interested in digital preservation, including NARA, but they may or may not be required by NARA to preserve records. Here a modest level of funding in a jointly funded program of research looking broadly at digital preservation problems might be appropriate.
- *How to preserve Web sites.* Government Web sites publish information in a wide variety of formats, use static and dynamic information, use information from a single source or many sources, and present general or user-specific views of information. Government Web sites also collect information from the public—for example, public comments on proposed regulations. What government Web site material should be archived? The variety of problems posed by government Web sites could be viewed as daunting, but it can also be an advantage in that the sites afford a means to gather much valuable information. Of course not all Web sites produce unique "records"; some are merely access portals to an underlying set of records that may be archived in other ways. Web sites should be a particular priority for archiving because they are so dynamic and also because they provide information that may not be captured by other archiving processes. An essential part of the national historical record may have simply been lost because government Web sites were not systematically archived in the past decade. (Serious efforts aimed at archiving Web content, however, are being made by such organizations as the Internet Archive.) NARA will need to co-develop policies and technical approaches for handling Web sites.
- *The interplay between information technology and human and organizational behavior related to records.* NARA, like many other organizations, has an interest in understanding how people and organizations use new technologies and an interest in the implications of this use for new forms of communication, new types of records, and new divisions of responsibilities for administrative and record-keeping activities. Relevant areas of research include business-process redesign, organizational restructuring as a result of new information technologies, adoption and adaptation of new technologies, work flow analysis and design, compliance with policies and requirements, and design of incentive mechanisms.

Problem Areas in Which NARA and Similar Institutions Have Special Interests

- *How to tease out (or define) records in interactive systems.* Interactions with and within government increasingly involve ephemeral views not stored as records but merely computed from online databases through software that itself changes over time. Web sites (discussed above) are one important example.
- *Transactions systems.* How should government transactions systems be preserved? Are there uniform ways (e.g., logs or snapshots) to archive them? Or should transactions systems be designed to emit records in order to preserve records as part of the standard work flow?
- *Providing digital assurances over a long period of time.* The computer science and cryptogra-

phy to support authenticity, integrity, and chain-of-custody requirements associated with long-term preservation are likely to be leading edge for the foreseeable future. A research program would explore new techniques to be able to assure integrity and authenticity of records held for many decades. Is an end-to-end scheme that would span hundreds of years possible?

CONCLUDING REMARKS

With necessarily limited resources, a limited capacity to manage research programs, and a large set of research challenges related to the general topic of archiving and digital preservation, NARA has to decide how and where it should invest. Questions for NARA to consider include these: What agencies share an interest in NARA's problem or a similar problem? What research community is working on that problem or a similar problem? How can NARA most effectively join with and influence that community? Does NARA need to sponsor its own research, or can it learn from or participate in others' research programs? What results are anticipated, and how will the results be brought back to NARA?

As the ERA starts to operate, NARA will surely uncover new problems. Some of the problems will lead to evolution of the system and its operation in straightforward ways—that is, known methods or engineering techniques will solve them. But operation may also reveal deeper problems worthy of new research activities.

5

Record Integrity and Authenticity

An archive is vitally concerned with the authenticity and integrity of its holdings. An "authentic" record is, loosely, one that is what it purports to be: it was duly issued by an authorized person or agency. A record has "integrity" if it is preserved without any alteration that would impair its use as an authentic record.

The issue of *digital* record assurance, the topic of this chapter, is discussed in greater detail than are other technical issues in this report for two reasons. First, the committee did not address this topic in depth in its first report, which emphasized other system design and related acquisition issues. Second, digital assurance is an area in which correct implementation requires great care.

To ensure the authenticity and integrity of paper records, archivists have worked out techniques such as establishing a chain of custody from a record's issuer to a record's user. Although the same techniques can be applied to digital records, the properties of digital records compel the use of additional techniques for ensuring authenticity and integrity. Digital techniques are available that can provide much stronger assurances than can existing techniques for paper records. Also, digital records are potentially more vulnerable to forgery and tampering—an attacker with access to the archive's computer could add, delete, or alter records in a wholesale fashion or make subtle alterations that would be difficult to detect by inspection. Increased experience with personal computers has made the public aware of how easily digital documents can be altered undetectably. The committee firmly believes that within the decade, both the public and the courts will have little confidence in digital records that lack the best assurances that technology can provide.

Digital checks have another huge advantage: an archive can verify the integrity of all records periodically. If errors are found—whether due to hardware or media failures, operator errors, or unauthorized changes—a redundant, undamaged copy of the record can be retrieved. When this form of auditing is performed often enough, errors in the archive can be fixed before they spread (e.g., before a redundant copy is created from an erroneous record).

A comprehensive system design that addresses authenticity and integrity has many pieces and a great many design and operational details. This chapter first describes basic tools and principles for digital record assurance and then offers some basic detailed approaches. Those approaches are presented as examples of the level of care that must be applied; it is not claimed that they are the only possible approaches.

DIGITAL ASSURANCE TOOLS AND PRINCIPLES

Digital assurances for records are based fundamentally on maintaining multiple, geographically and administratively separated copies and on using cryptographic techniques to provide integrity checking and secure transmission of records to and from the archive. These techniques and their appropriate application to a long-term archive are discussed below.

Geographical Replication

Multiple, geographically and administratively separated replication provides an essential technique for protecting integrity. There are various ways to meet this requirement, involving complete replicas or multiple partial replicas. A detailed design starts with reasonable goals for the acceptable bit loss rates and desired availability for each local site. Given these reliability metrics for the local sites, one can compute the reliability of an *N*-way replicated archive, and the total archive can be designed to achieve a specified level of reliability.

At any given point in time, the engineering and design question will be how to achieve these goals in a cost-effective manner with currently available storage technologies. Storage technologies and their relative prices are evolving rapidly, so the National Archives and Records Administration (NARA) needs a flexible design that can evolve along with them.

Cryptographic Techniques

As described below, cryptographic techniques provide basic tools for ensuring authenticity and integrity.[1] Both qualities depend on cryptographic algorithms for which forgery is computationally infeasible: that is, defeating the system (e.g., by altering the original record without altering the digest or its signature) would require so many samples of records or so much computing power that no attacker has the resources to succeed.

Integrity Check Using Hash Digests

The technique of computing a hash digest or checksum of a record is used to check its integrity. A secure hash algorithm computes a compact hash digest from the digital bits that comprise a record. There are several algorithms in common use; Federal Information Processing Standards (FIPS) Publication 180-2[2] specifies four standard algorithms. The standard explains that a secure hash algorithm is one for which "it is computationally infeasible (1) to find

[1]More details on these and other assurance techniques can be found in the following reports from the National Research Council's Computer Science and Telecommunications Board: *Cryptography's Role in Securing the Information Society* (1996), *Computers at Risk*: *Safe Computing in the Information Age* (1991), and *Trust in Cyberspace* (1999), all published by National Academy Press, Washington, D.C.

a message that corresponds to a given message digest, or (2) to find two different messages that produce the same message digest. Any change to a message will, with a very high probability, result in a different message digest."[3]

In the field of databases, "witness" is the technical term for a hash digest that is computed when a record first enters an integrity-management system. One later verifies the integrity of an offered record by computing a new hash digest from the offered record and comparing that new value with the witness.

Secure Transmission

Various cryptographic techniques can be used to ensure that what is received is what was sent. A digital record can be authenticated by creating a digital signature that can be verified by anyone using the record. If the record is altered in any way, the signature check will fail. To sign a digital object, one performs a key-driven cryptographic transformation, typically on a witness of the digital object (such as a hash digest) rather than on the object itself, to create what is known as an authentication tag or signature for the given digital object. To verify the origin of a digital object, one performs an algorithm that involves a second key-driven cryptographic transformation, taking as input the authentication tag and a public key of verified authenticity; the output is a binary value, "origin verified" or "origin not verified." The meaning of "origin verified" is that the object originated with someone who possessed the private key. One must trust that the putative possessor of the private key has properly protected it, so that he or she is the only one who could have created the authentication tag.

It is important to recognize that digital signatures have limited value for the long-term preservation of a chain of custody or for data integrity. Their value is limited by their validity window: the time-to-compromise of the secret signing key, the time-to-compromise of the signature algorithm, and the time-to-obsolescence of the public key infrastructure—whichever is shorter. For example, if a private key used to form digital signatures for records becomes compromised as of a certain date, any records verified with the public key corresponding to that private key after that date are suspect. The case of a private key becoming compromised after it is no longer in use is more subtle: while the compromise does not endanger records already verified, it allows the attacker to forge a document and to record its creation date as one during which the key was still valid. Discovery of the compromise of a key may come a long time after it actually occurred. Compromise could also occur as a result of a cryptoanalytic attack.[4] Digital signatures are, therefore, an excellent means of verifying that recently transmitted data—such as a set of records being transferred from an agency to

[2]National Institute of Standards and Technology (NIST). 2002. Federal Information Processing Standards [FIPS] Publication 180-2. Information Technology Laboratory, NIST, U.S. Department of Commerce, Washington, D.C., hereafter referred to as FIPS Publication 180-2. Available online at <http://csrc.nist.gov/publications/>. Accessed May 23, 2005.

[3]NIST. 2002. FIPS Publication 180-2.

[4]The first widely distributed description of the RSA algorithm for public-key encryption was in Martin Gardner, 1977, "Mathematical Games: A New Kind of Cipher That Would Take Millions of Years to Break," *Scientific American* 237(2):120-124. The article included a challenge ciphertext, which was decoded in 1994, as reported in Derek Atkins, Michael Graff, Arjen K. Lenstra, and Paul C. Leyland, 1995, "The Magic Words Are Squeamish Ossifrage," *Advances in Cryptology (ASIACRYPT'94), Lecture Notes in Computer Science* 917:265-277.

NARA—actually came from where it is claimed that they came from, and a poor means for authenticating the far-in-the-past origin of stored data.

Digital signatures can be used together with several other cryptographic tools to establish a secure communications channel. The Secure Sockets Layer (SSL) protocol, first implemented in Netscape browsers, later standardized by the Internet Engineering Task Force, and implemented today in all widely used Web browsers, is an example of a secure channel. It establishes a secure channel on top of the insecure channel provided by the Internet's Transmission Control Protocol (TCP) by an exchange of credentials (using digital signatures), negotiation of shared secret keys, and then exchanges of message authentication codes (MACs). The MACs are essentially a shared-secret-key signature mechanism. One important step that is sometimes overlooked in the application of SSL is that the protocol reports who is at the other end of the secure connection. For authentication to be complete, the recipient must check that report to see that the correspondent that it identifies is the expected one. Many current Web browsers that use SSL either omit or deemphasize this step, with the consequence that their supposedly secure channel may not be secure at all.

Long-term record assurance depends on measures other than digital signatures or secure channels. It requires (1) at record ingest, securely transmitting the data from its origin to NARA and recording metadata about that assurance; and (2) from then on, maintaining the integrity of the stored data and associated metadata while the record resides in the archive. These two steps, ingest and retention, are examined in the following sections.

ASSURANCE AT RECORD INGEST

Verification of Validity

Agencies that create and retain records must use appropriate techniques for authenticating the origin of records and for maintaining their integrity while they remain in the custody of the agency. When a record is transferred to NARA, the agency should transmit, along with the record, metadata that exhibit an audit of the record's assurances using the agency's internal techniques. If the agency maintained integrity checks, these should be included as part of the metadata.

Digital signatures may be helpful in this verification, depending on the particular circumstances. If a set of records were digitally signed and the digital signature system is still operational and the validity window (see above) is believed not to have expired, agencies can provide additional information about record validity by checking the signatures and recording the result in metadata. On the other hand, it would be problematic for NARA itself to perform this verification, because the verification depends on the operation of a system outside NARA's control. Moreover, if an agency retains custody of its records for a lengthy period, it must face the prospect that the validity window of its digital assurances expires before the records are transferred to NARA.

Once a record's authenticity has been verified and the record ingested, the primary requirement from then on is to record the verification method and outcome in the metadata and to use integrity assurances to protect the integrity of both the record and the metadata while they are retained in the archive.

Assurance in Transmission

As NARA evolves away from transfer and storage based on the delivery of physical media toward network-based transfers and online storage, the use of more robust safeguards in transferring records from agencies to NARA becomes even more important. For example, verifying that an electronic file transfer via File Transfer Protocol (FTP) yields a file with the right number of bytes provides little assurance against either unintentional corruption or deliberate tampering.

Assurances for the transmission between the Electronic Records Archives's (ERA's) ingest component and the originating agency's system are needed to provide adequate assurance that the transmitted data come from the claimed source and thus that an unauthorized person has not attempted to submit false records and that the records have not been altered in transmission. The best option is for NARA to use a standard and well-vetted scheme rather than a custom scheme (which would necessarily have been subject to less scrutiny).

A good example of such a standard, widely used scheme today is a properly authenticated SSL connection. Records themselves can be sent securely via such a channel. Alternatively, a secure channel can be used to send individual hash values for the records, and the records can be sent any which way as long as they are checked against the hash values at NARA's end. If Standard Form 258 ("Agreement to Transfer Records to the National Archives of the United States") becomes electronic in the future, it would also be appropriate to authenticate and securely transmit these forms as well.

Guarding Against Human Error

Multiple copies and integrity checks can protect against mechanical threats to integrity, but they do not address the biggest remaining threat to integrity: human error. To illustrate: a careful look at "agency report 493" reveals it to be another copy of "agency report 492," accidentally created when someone clicked on "Revert" just before clicking on "Save as." Such errors can occur within an agency before the records reach NARA or during the ingestion process.

NARA ingest processes should be designed to minimize the opportunity for human error of this sort. In addition, NARA may be able to afford to have humans occasionally inspect the incoming records and query originating agencies to resolve problems. The results of such interactions with agencies would constitute additional metadata (e.g., "at ingest, so-and-so received assurances from so-and-so that this record, and not record xxyyxx999-04444, is the authentic representation of the document with the given title"). In other words, ingest negotiations add additional metadata. This observation is not intended to suggest that NARA undertake comprehensive examination and repair of records at ingest.

ASSURANCE DURING RETENTION

The starting point of a full design for the assurance of archived records is to carry out a careful, comprehensive threat assessment (see the discussion below). This section considers several of the obvious threats and techniques available to address them.

Designing digital assurances into an electronic records archive is similar to designing security measures. First, the cryptographic techniques must be chosen carefully, and the

archive must be ready to change the cryptographic algorithms that it uses. For example, if someone were to discover a way to create two different documents that generate the same hash digest with the current hash algorithm, the algorithm would need to be replaced. Second, the overall system design, not just the cryptographic mechanisms, must not allow openings for attackers. For example, if an attacker can surreptitiously tinker with the code that NARA runs to perform signature or hash code verification (to make it say "yes" when it should have said "no"), the assurances are worthless. If such vulnerabilities are discovered, the system will require modification.

Cryptographic Integrity Assurance

Multiple, geographically separated copies provide the fundamental basis for establishing integrity. Hash algorithms provide a performance enhancement for checking that records have not been corrupted, whether by hardware failure, system errors, or deliberate tampering. The integrity of a record can be verified by computing the hash digest for the bits in the record and comparing this digest to a previously calculated and separately protected witness for the record. If the two digests match, the integrity of the record is established to a very high degree of probability. A witness computed for a bit stream that includes the record plus metadata representing the time at which the hash was computed provides a digital time stamp that allows one to verify the contents of an object at the time the witness was calculated and published.

Although the description of integrity checking presented here has assumed that each record carries a separate hash digest, this need not be the case. After all, integrity considerations apply not only to individual records but also to collections, or series, of records. A town clerk who issues 10 marriage licenses on July 12 passes 10 records to the archive, but also attests that these records constitute the entire collection of licenses issued on that date—that is, that these and only these licenses were issued on July 12. A hash can be computed for a set of records in a given order. Checking such a hash may require retrieving all of the records on which it is based and recomputing the hash, which might be time-consuming. In exchange, however, fewer known good hashes have to be published.

To protect against malevolent change by an attacker, a record's witness must be separately protected so that an attacker who manages to gain access in order to change the record cannot also alter the witness. Since a hash digest can be written in a relatively small number of digits, one way to protect it from change is to publish it in a very public place, such as a classified advertisement in a major newspaper (which will, a short time after publication, be distributed on microfilm to many hundreds of libraries), or by otherwise depositing the hash value in hundreds of libraries. (Such a hash publication service could be provided by the Government Printing Office, which currently distributes government documents to Federal Depository Libraries.) Because integrity checking requires being able to establish a relationship between a record and its published hash digest, a unique record identifier associated with each record is very useful. The published list is then simply a correspondence between record identifiers and hash digests.

Techniques for combining hash values can be used to reflect the integrity of a huge archive using a single hash digest, which is more practical to publish than a very large number of individual hashes. The most straightforward approach would be to compute the single hash digest formed by hashing the hash digests of every record in the archive, but such a sequential computation would be prohibitively time-consuming. Also, a too-simple combination such as concatenating all hash values may lower the precision of the integrity check. Techniques have

been devised that require far less computation to update a single hash value when the archive is changed and do not lower the precision of the check.[5] The important property of all techniques that compute a single hash value is that the value irrefutably depends on the precise bit-string contents of every record in the archive. [6]

Other Assurance Measures

Cryptographic protections are only a part of the solution, however. They detect damage or attack only after the fact. Prevention and repair are even more crucial, and they depend on careful procedures and system designs for handling digital records. They include the maintenance of multiple, geographically and administratively separated copies; access controls; audit logs; retention of old or deleted files to recover from unauthorized or faulty operations; and procedural safeguards for changing the archive contents to reduce human error (e.g., requiring multiple people to authorize a change). These protection measures are important in order to avoid or correct tampering with the archive. Although cryptographic techniques can detect that tampering has occurred, they are not able to deter or repair tampering.

Contingencies

What if stored records are found to contain errors? What if one or more of the cryptographic algorithms on which digital assurances are based are shown to be easy to subvert? Over the life of records in the archive, some of these events are sure to occur. Thus, the archive must be designed to anticipate corrective responses.

Correcting Post-Ingestion Errors

Errors may creep into the archive after records have been verified and ingested. This could happen as a result of hardware failure, operational error, a software bug, malicious attack, or something else. Errors are detected by comparing each file in the archive with another copy or by reading the file, calculating its current hash, and comparing that hash with the separately protected witness for that file. There must be an ongoing process that performs such comparisons so that errors do not accumulate unnoticed.

---

[5]One such approach, first suggested by Ralph Merkle (1980, "Protocols for Public Key Cryptosystems," pp. 122-133 in *Proceedings of the 1980 Symposium on Security and Privacy*, IEEE Computer Society Press, Los Alamitos, Calif.), is to build a binary tree out of the million witness values, as follows. The leaves are these values, and each internal node is the hash of the concatenation of its two children. The root of this tree is published so that it is widely available and hard to tamper with, such as in a major newspaper that is stored on microfilm in libraries across the nation. Stored with each of the million witness values is the following linking information: the list of 20 sibling hash values (each one accompanied by a bit indicating whether it is the right or the left sibling) along the path from the leaf up to the published hash value.

[6]The Merkle-tree technique was adapted for the purpose of digital time-stamping by D. Bayer, S. Haber, and W.S. Stornetta, 1993, "Improving the Efficiency and Reliability of Digital Time-Stamping," pp. 329-334 in *Sequences II: Methods in Communication, Security, and Computer Science*, R.M. Capocelli, A. De Santis, and U. Vaccaro (eds.), Springer-Verlag, New York, N.Y., and by Josh Benaloh and Michael de Mare, 1991, *Efficient Broadcast Time-Stamping*, Technical Report No. TR-MCS-91-1, Clarkson University Department of Mathematics and Computer Science, Potsdam, N.Y. The technology is offered commercially by Surety, Inc.

When integrity checks fail, both repair and investigation are required. First, invalid data must be restored from mirrors or backups provided by redundancy mechanisms designed into the archive. Second, the cause of the error must be investigated. Failure to rigorously investigate cases of failed integrity will eventually lead to failure to protect the archive; modifying the system design in response is an essential part of iterative design.

If it turns out that files have been improperly altered by the ERA software or its operators, it will be necessary to "undo" certain changes to the archive. Repair can be facilitated by designing the file system to allow operations that modify files to be "undone," at least to some level; techniques similar to those used to implement "undo" in desktop word-processing software can be applied to a file system as well. The most effective technique will be to limit the amount of the archive that can be modified at all: ideally, most storage would be treated as "read only" by the software; only portions of the archive in which collections of records were being assembled during the ingest process would be modifiable. Read-only records may, of course, be subject to data-type transformations so as to combat obsolescence, but these processes should only add new forms, not delete the old forms.

Compromise of Cryptographic Algorithms over Time

Cryptographic algorithm compromise is a distinct possibility that must be assumed to happen at some point during the lifetime of the ERA. Breakthroughs in mathematics, cryptanalysis, or computing power over the lifetime of the ERA may result in successful attacks on the one-way functions that form the basis of digital assurances used in the archive. For example, results reported at the 24th Annual International Cryptology Conference (held in Santa Barbara, Calif., in 2004) indicated some weaknesses in hash functions commonly used today. After the conference, NIST announced that it plans to phase out the use of Secure Hash Algorithm (SHA)-1, the recommended federal standard cryptographic hash function, by 2010. In mid-February 2005, it was announced that a team of cryptanalysts at universities in China had invented a new hash-collision algorithm attacking SHA-1 that is substantially faster than all previously known attacks. While these results are far short of a devastating compromise of existing standard algorithms, they remind us that the algorithms may not be impregnable.

If a cryptographic algorithm were to be compromised, hash values computed using that algorithm would not provide assurances of integrity. The following provisions are, therefore, essential for long-term archival preservation:

- Archive systems should be designed to accommodate replacement of cryptographic functions and to allow records to be reprocessed to attach revised digital assurances. Well-designed security and signature systems accommodate more than one cryptographic algorithm for just this reason; such provisions are especially important owing to the very long lifetime of an archive system.
- Replacement of the algorithm and recalculation of hash values should be performed both when compromise is threatened and when more-robust hash algorithms are developed and accepted by the cryptographic community.
- Better algorithms should be adopted early rather than waiting until the older algorithms are known to have been compromised. This action would guard against the old integrity schemes being compromised, unbeknownst to NARA, before the new one was applied.

BOX 5.1
How to Introduce a New Integrity Scheme into an Archive

One way of introducing a new integrity scheme would require that, in addition to the hashing system for integrity, there would also be a secure time-stamping system. Suppose that an implementation of a particular time-stamping system is in place, and consider the pair (r, c_1), where "c_1" is a valid time-stamp certificate[1] (in this implementation) for the digital record "r." Now suppose that some time later an improved time-stamping system is implemented and deployed—by replacing the hash function used in the original system with a new hash function, or even perhaps after the invention of a completely new algorithm. Is there any way to use the new time-stamping system to buttress the guarantee of integrity supplied by the certificate c_1, in the face of potential later attacks on the old system?

One could simply submit r as a request to the new time-stamping system, but this would lose the connection to the original time of certification. Another possibility is to submit c_1 as a request to the new time-stamping system, but that would be vulnerable to the later occurrence of a devastating attack on the hash function used in the computation of c_1, as follows: if an adversary could find another record r' with the same hash value as r (a hash collision), then the renewal system could be used to backdate r' to the original time.

Suppose instead that the pair (r, c_1) is time-stamped by the new system, resulting in a new certificate "c_2," and that some time after this is done (i.e., at a definite later date), the original method is compromised. The certificate c_2 provides evidence not only that the record contents r existed prior to the time of the new time-stamp, but also that it existed at the time stated in the original certificate, c_1; prior to the compromise of the old implementation, the only way to create a valid time-stamp certificate was by legitimate means.[2]

[1]Whatever output a time-stamping system produces in response to a time-stamp request for a particular bit string, such as a signature returned by a hash-and-sign system.

[2]This approach first appeared in the technical literature in D. Bayer, S. Haber, and W.S. Stornetta, 1993, "Improving the Efficiency and Reliability of Digital Time-Stamping," pp. 329-334 in *Sequences II: Methods in Communication, Security, and Computer Science*, R.M. Capocelli, A. De Santis, and U. Vaccaro (eds.), Springer-Verlag. The paper is available online at <http://www.surety.com/solutions/DN/bhspap.pdf>, accessed May 1, 2005.

• Archive systems should also support multiple algorithms simultaneously. It would be prudent to use two hash functions in parallel in anticipation of future cryptanalytic advances. In addition, migration to new hash schemes would likely be performed over time, as records are copied from one medium or storage system to another, rather than all at once.

Box 5.1 describes an approach to putting a new integrity scheme in place.

The issue of algorithm compromise will be an ongoing operational concern. Someone must pay enough attention to keep track of the need to apply a new integrity scheme and must make a judgment call about when to add the new scheme to the system. Once the new scheme is added, new hash values can readily be calculated during the course of routine scans of the archive for errors.

By using standard schemes, NARA can leverage the knowledge of a wide community of cryptographers and system designers to inform its decisions over time. For example, NARA can look to NIST, which monitors developments in this area and issues advisories, for advice.

ASSURANCE AT DELIVERY TO ARCHIVE USERS

When it delivers an electronic record or batch of records to a customer, NARA can use a secure transmission method or a digital signature, which provides assurance that the data can be trusted because they originated at NARA (which checked the origin of the data at ingest and then held the data in a trusted repository). Such protections also provide assurance that data have not been tampered with in transit after leaving the ERA system.

ASSURANCE FOR ADDITIONAL INFORMATION

The records themselves, in their original forms, are considered above. But additional information, such as derived forms and metadata, should also be protected with the same level of assurance.

Assurance for Record Migration and Derived Forms

During its operation, the ERA may need to create alternative forms of the records that it ingests. These alternatives may be needed to simplify meeting ERA users' needs (derived forms, as described in the committee's first report), to prevent records from becoming unusable if the data type in which they are encoded becomes obsolete, or to meet the needs of certain users (e.g., for the redaction of classified records). In any case, the new form of the record will not contain the same bits as those in the original, and as a result neither the authentication nor the integrity checking associated with the original record will apply to the new form. Nevertheless, a chain of trust can be established from the original record to the new form, assured by cryptographic techniques.

The basic idea is that the operation that transforms a record in order to create a new one must certify the following: (1) the record it used as input, (2) the authority imputed to the transformation process, and (3) the integrity of the output of the process. Whether the transformation involves only computer processing or includes manual processes (e.g., redaction), the transformation process must make such a certification. It includes an unambiguous identification of the input record (e.g., record identifier and hash digest), an unambiguous identification of the output record (e.g., record identifier and hash digest), and perhaps other information such as the identity of the person performing the redaction or the version number of the software that performs a data-type conversion. The transformation process or processes are thus unambiguously identified, and the user of the record in its new form is thus presented with an assurance history of the document. Users can then decide, based on the trust they impute to any transformation processes applied, whether to trust the new record.

The scientific community that records, processes, and analyzes sensor data is grappling with similar problems. Agencies and disciplines seeking to keep track of these transformations applied to streams of data include NOAA (weather), NASA (satellite imagery), and high-energy physics (records of particle accelerator experiments). Developments in these areas may prove useful to NARA in the future.

Assurances for Metadata

Some metadata is an essential accompaniment to a record and must be protected accordingly. For example, the date on which a record was created may not be a part of the record, but is instead recorded as metadata. This metadata must be protected by digital assurances, just as is the rest of the record.

Some metadata may change over time, being revised as metadata standards change or as better techniques become available for extracting metadata from record contents. This information is similar to a derived form of a document—that is, it should be assured by an audit trail that identifies the original input record and the processes used to derive the new metadata.

THREAT MODELING AND THREAT COUNTERING

Threat modeling is a systematic technique for enumerating threats against digital systems. It includes, among other elements, developing scenarios involving accidental or deliberate misuse of systems that might lead to vulnerabilities. The model is then used to analyze a system design to understand how it performs in different scenarios.

After developing a threat model, designing ways to counter the expected threats, and building the system, one should then observe the system to see what kinds of integrity, authenticity, and security failures occur despite carefully laid plans. Next, this failure analysis is used to adjust the threat model, to design additional or better threat-countering procedures, and to upgrade the system. The process is repeated for the entire lifetime of the system. Systems that are to provide fault tolerance, integrity, security, or life safety all require this kind of continual iteration throughout their entire life cycle. The buzzword is "design for iteration," and this style of system design makes the feedback mechanism an essential component of the system design.

EVOLUTION OF ASSURANCE OF RECORDS

An ideal scheme for digital assurance would afford an end-to-end verification of authenticity and integrity. That is, the reader of a record, many decades after its creation, would be able to directly verify the authenticity of the creator and the integrity of the record's preservation. As described above, current techniques are limited to piecemeal assurances that together bridge the validity windows of key management systems, of cryptographic algorithms, and data-type transformations. An ideal scheme is not currently available; perhaps research can improve on the techniques that we depend on today (see Chapter 4).

With present techniques, the chain-of-custody and assurance bridging techniques are only as strong as their weakest link. An archive that superbly guarantees the integrity of its records will not be useful if the agencies sending records to the archive have been sloppy about any aspect of stewardship of the records in their custody. To ensure that the ERA, decades hence, will be valued as a source of authentic government records, the entire chain of record custody must use robust assurances.

Promoting Digital Assurance Throughout the Federal Government

The preceding arguments clearly call for digital assurances to be applied to records throughout their life, not just starting at the time they are ingested by NARA. However, digital

assurance techniques are not currently widespread in either government or commercial information technology (IT) systems or electronic records-management systems.

NARA's mission to preserve essential evidence suggests that it take a much more active role in promoting digital authentication and integrity assurance techniques that agencies can use to safeguard the records they create. If NARA cannot provide state-of-the-art attestations about the records it holds, the records will not be honored as valid in an environment in which significantly better practices exist.

A well-designed and well-operated ERA itself can serve as an exhibit of "best practices" in the digital retention of electronic records, including both technical and operational methods for assuring record authenticity and integrity. Also, as new government IT systems are developed and as new records-management systems are developed or procured, NARA should assist agencies in their adoption of acceptable digital assurance measures. These techniques require software to create and maintain digital signatures and the like, but also operational measures to issue and manage cryptographic keys, and operational measures to ensure that the records-management system itself is not compromised.

The Records Management Redesign initiative, currently underway, provides an opportunity to inaugurate and promote the use of digital assurance techniques for government records. The thrust of this initiative is to engage the record-creating agencies in the overall records-preservation mission, taking on responsibility for defining, creating, and maintaining digital records in a form that streamlines the preservation and later use of the records. Strong assurance is a vital aspect of electronic records preservation that has not received adequate attention.

Until record creators use digital assurance methods that conform to NARA's standards, agencies that create and hold electronic records should be held to stringent chain-of-custody standards for their holdings. When records without digital assurances are transferred to NARA's custody, NARA should immediately augment these records with suitable digital assurances, which can then be maintained throughout the records' life in the ERA. The ultimate goal must be to achieve and maintain the best digital assurances and other record-retention practices in both the creating agencies and the archive itself.

Appendixes

Appendix A

Briefers to the Study Committee

Although the briefers listed below provided much useful information of various kinds to the committee, they were not asked to endorse this report's conclusions or recommendations, nor did they see the final draft of the report before its release.

AUGUST 8-9, 2002
WASHINGTON, D.C.

John Carlin, Archivist of the United States
Robert Chadduck, Research Director, Electronic Records Archives Program, National Archives and Records Administration (NARA)
Michael Lesk, National Science Foundation
Reagan Moore, San Diego Supercomputer Center
Kenneth Thibodeau, Director, Electronic Records Archives Program

NOVEMBER 4, 2002
NATIONAL ARCHIVES AND RECORDS ADMINISTRATION
COLLEGE PARK, MARYLAND

NARA Staff

Margaret Adams, Electronic and Special Media Records Services Division
Nancy Allard, Co-director of Electronic Records Management (ERM) Initiative, Policy and Communications Staff
Bruce Ambacher, Electronic and Special Media Records Services Division
Michael Carlson, Director, Electronic and Special Media Records Services

Robert Chadduck, Research Director, Electronic Records Archives Program
Allen Church, Electronic Records Archives Program Office
Susan Cummings, Archives Specialist, Policy and Communications Staff
Fynnette Eaton, Electronic and Special Media Records Services Division
Sharon Fawcett, Office of Presidential Libraries
William Harris, Office of Presidential Libraries
Mark Huber, Electronic Records Archives Program
Greg Lamotta, Electronic and Special Media Records Services Division
Dyung Van Le, ERA Systems Engineering Director
Edwin McCeney, Federal Records Officer for the Department of the Interior
Richard Steinbacher, Electronic Records Archives Program
Kenneth Thibodeau, Director, Electronic Records Archives Program
Steven Tilley, Chief of the Special Access/Freedom of Information Act (FOIA) Staff
Haseen Uddin, Chief Technology Officer
Jon Valett, Electronic Records Archives Program
Sam Watkins, Office of Human Resources and Information Services
Lisa Weber, Information Resources Policy and Administration Division
Yvonne Wilson, Office of Records Services

NOVEMBER 5, 2002
WASHINGTON, D.C.

James Gray, Microsoft Research (by telephone)
Joseph King, National Space Science Data Center
Michael Miller, Archivist, Federal Bureau of Investigation
Anna Karlin Nelson, American University
Jim Olson, Sabbath, Inc. (by telephone)
James Ostell, Chief, Information Engineering Branch, National Center for Biotechnology Information
Anne Van Camp, Research Libraries Group

FEBRUARY 27-28, 2003
WASHINGTON, D.C.

Rick Barry, Principal, Barry Associates
Laura Campbell, Associate Librarian for Strategic Initiatives and Chief Information Officer, Library of Congress
MacKenzie Smith, Associate Director for Technology at the MIT Libraries, Massachusetts Institute of Technology

MAY 8, 2003
WASHINGTON, D.C.

Lewis Berman, Branch Chief of the Informatics Branch of the Division of Health Examination Statistics, Centers for Disease Control and Prevention (CDC)

Lee Dantzler, Director of the National Oceanographic Data Center, National Oceanic and Atmospheric Administration
James Erwin, Director of Information Science and Technology, Defense Technical Information Center (DTIC)
Chris Olsen, Chief of Records and Classification Group in the Office of the Chief Information Officer, Central Intelligence Agency
Charles Rothwell, Associate Director for Information Technology and Services of the National Center for Health Statistics, CDC
Tim Sprehe, President of Sprehe Information Management Associates, formerly with the Office of Management and Budget and the Census Bureau
Pete Suthard, Director of Operations, DTIC
Jeanne Young, Federal Reserve Board (retired)

MAY 29, 2003
SITE VISIT TO GOOGLE, INC.
MOUNTAIN VIEW, CALIFORNIA

Howard Gobioff, Google, Inc.
Brian Rakowski, Google, Inc.

AUGUST 14, 2003
NATIONAL ARCHIVES AND RECORDS ADMINISTRATION
COLLEGE PARK, MARYLAND

NARA Staff

Nancy Allard, Co-director of ERM Initiative, Policy and Communications Staff
L. Reynolds Cahoon, Assistant Archivist for Human Resources and Information Services and Chief Information Officer
Robert Chadduck, Research Director, Electronic Records Archives Program
Susan Cummings, Archives Specialist, Policy and Communications Staff
Mark Giguere, Co-director of ERM Initiative, Modern Records Programs
Michael J. Kurtz, Assistant Archivist, Office of Records Services
James McKan, Executive Officer, Electronic Records Archives Program
Kenneth Thibodeau, Director, Electronic Records Archives Program

Appendix B

"Summary and Recommendations" Chapter from the Committee's First Report

As in other sectors of society, much of the business—and thus record keeping—of the federal government depends on digital information. Documents are created, transmitted, and stored electronically. E-mail has become an important—and often primary—communications technology. And many records exist only in electronic form, stored in databases and other computer systems. Some of these records will, in time, be transferred to the custody of the National Archives and Records Administration (NARA) for long-term preservation.

NARA's current systems for archival preservation of electronic records are limited in capability and ad hoc in nature. Recognizing the growing importance of electronic records to its mission of preserving "essential evidence,"[1] NARA launched the Electronic Records Archives (ERA) initiative. It sponsored work through this program at the San Diego Supercomputer Center (SDSC), which resulted in a series of archival preservation demonstrations. Building on this experience and that of other institutions studying digital preservation, NARA's new ERA Program Office plans to begin initial procurement for a production ERA in 2003. As of this writing, NARA has hired a contractor to assist with the ERA program and has started to define desired capabilities and requirements for the system, including a vision statement and concept of operations for the ERA.

NOTE: Reprinted from National Research Council, 2003, *Building an Electronic Records Archive at the National Archives and Records Administration: Recommendations for Initial Development*, Robert F. Sproull and Jon Eisenberg (eds.), The National Academies Press, Washington, D.C.

THE IMPORTANCE OF THE ELECTRONIC RECORDS ARCHIVES PROGRAM (CHAPTER 1)

Finding 1. As NARA recognizes, it is critical to start developing new electronic records preservation capabilities quickly in order to continue to fulfill NARA's mandate to preserve federal records.

With the rapid increase in federal records in digital form—and with many records born digital or existing only in digital form—it is clear that solutions must be found for preserving these records in order for NARA to continue to fulfill its mandate. NARA has determined and the committee concurs that new capabilities for electronic record archiving are needed for NARA to perform its mission to preserve and provide access to federal records of enduring value.

The overall challenge facing NARA is substantial. The volume and diversity of digital records that will be eligible for transfer from the custody of federal agencies to NARA are projected to be very large. Indeed, it is reasonable to anticipate that in the not-too-distant future, the number of digital records is likely to exceed the number of records originating in paper form. NARA's current systems for electronic records, designed primarily to support preservation of relational databases and similar highly structured records, cannot meet these demands.

The backlog of electronic records presents additional challenges. Under the paper-based model, NARA receives records from a few years to many decades after they were are created. The transfer of electronic records to NARA has for the most part proceeded in similar fashion. Thus when the ERA system becomes operational, NARA will face a large backlog of electronic records that were created over the past few decades, many of which may pose challenging preservation problems owing to their age (media deterioration, loss of documentation and other metadata, and obsolescence of data types). For records yet to be created, there may be ways to avoid the technology obsolescence problem by restructuring records acquisition processes to obtain records closer to the time they are created. (Discussion of this opportunity and related process issues is deferred to the committee's second report.)

If NARA fails to design and implement an electronic records archiving program that is capable of handling the projected volume and diversity of electronic records, important records are likely to be lost for the reasons discussed above. Likewise, significant delays in the ERA program would put records at greater risk of loss. The consequence of either failure to institute a program or a significant delay in doing so would be the possible—indeed even likely—loss of an important part of the nation's history.

Finding 2. ERA systems can and should be built, but it is a challenging, leading-edge engineering undertaking, not a routine procurement.

Although no one has yet designed, built, or managed a production digital archives system on the scale that NARA envisions, the ERA program can be launched in a technically sound way. No off-the-shelf overall solution is available, but there are demonstrated solutions to many of the important system components the ERA will need, making it possible to start building ERA capabilities today. The projected scale and complexity of the ERA program mean that the task of designing, engineering, and evolving the system is a formidable challenge.

Recommendation 1. The ERA should comprise a series of interrelated systems that evolve over time to fulfill NARA's digital preservation needs.

The scope of the ERA in terms of time and function demands that the digital archives be thought of as a set of systems that evolve over time. Digital materials (records and associated metadata) will need to be preserved for a very long time—longer than the lifetime of any physical device or software component that is part of the ERA. This means that hardware and software will have to be replaced or modernized many times without disrupting the archive.

Because requirements for NARA's digital archives will change over time, the ERA systems must be designed to evolve gracefully. For example, new data types (or "file formats"[2]) will emerge and require modifications to parts of the system. Likewise, storage and other implementation technologies will evolve, necessitating the replacement or upgrading of parts of systems without disrupting the operation of the ERA as a whole. The volume of records to be stored will continue to increase, requiring a strategy for graceful scaling of storage and processing. New options for preservation will probably emerge and be incorporated into the ERA. User demands will also change over time, requiring modifications in the ways that records are located and accessed by users.

COMMONALITIES WITH OTHER DIGITAL PRESERVATION ACTIVITIES (CHAPTER 2)

Finding 3. The requirements of NARA's ERA program have much in common with those of other digital preservation systems.

Although NARA's statutory mandate to preserve federal records is unique, many organizations need to preserve digital objects and are taking steps to design and implement systems that address this need. They are developing architectures, techniques, experiments, pilot systems, and expertise relating to digital preservation. NARA has embraced some of these developments, such as the Open Archival Information System (OAIS) reference model for characterizing archival systems in terms of their ingest, storage, and access functions,[3] and is aware of activities in organizations such as the Library of Congress, the National Aeronautics and Space Administration (NASA), and the Online Computer Library Center (OCLC).

The technology to address the ERA largely overlaps the technology required to build other digital repositories. For example, NARA and others require robust long-term storage of bits, accommodation for many different data types, metadata standards and processing techniques, and flexible searching and access provisions. Many systems require a high degree of scalability and the ability to evolve their architecture and implementation. And a number of organizations, government agencies, and businesses also face the challenge of long-term preservation of large volumes of data. As it examines commonalities with others working on digital preservation, NARA should evaluate what is already available as commercial software.

There are a few areas where NARA faces requirements that are more stringent than those typical of other digital repositories. NARA's mandate to guarantee record authenticity is stronger than that of some but not all organizations. NARA must also be ready to preserve materials of very diverse types, even if they were created by obsolete systems or saved on obsolete storage media; by contrast, many digital libraries can simplify their challenge by accommodating only a limited set of contemporaneous or common data types. Indeed, the ERA will have to be capable of ingesting the full variety of data types used to create permanent records across the federal government, which will roughly correspond to the full variety of data types in use more broadly. Also, NARA must sometimes redact classified or restricted documents in order to produce versions that can be released to the public. While these and other requirements may be special to NARA's electronic preservation systems, the list of special requirements is modest compared with the list of requirements that the ERA shares with other digital preservation systems.

Recommendation 2. NARA should emphasize the ERA program's commonality with other digital preservation systems and engage with other programs and organizations wherever possible.

NARA would benefit from increased coordination with other federal entities (such as the Library of Congress or NASA), other institutions (such as OCLC, university libraries, archiving projects, and foreign national libraries and archives), and businesses that share common interests in digital preservation. Enhanced coordination should extend at least to increased information sharing as the various institutions move forward with digital repositories. It might extend to such activities as joint work on standards or best practices. However, the committee is not recommending coordinated or joint procurements, which could significantly complicate or delay the ERA program.

It is not enough to be aware of efforts by other institutions; NARA must engage them by becoming an active participant. Engagement means not only that NARA will have better access to the expertise and artifacts that these efforts develop but also that it can influence the research, development, and deployment agendas of these groups when it is appropriate and help build a larger community addressing engineering issues related to digital preservation.

NARA's long-term objective should be to increase the commonalities between its preservation systems and those of other digital repositories such as libraries, because this will (1) make it possible to share software and metadata standards with other institutions, (2) help stimulate the development of commercial off-the-shelf (COTS) components by increasing the size of the market for these components, and (3) help grow the cadre of professionals trained in digital preservation and build ties that would help NARA recruit them to work on the ERA program.

LESSONS LEARNED FROM THE SDSC DEMONSTRATIONS (CHAPTER 3)

Finding 4. Demonstrations conducted at the San Diego Supercomputer Center (SDSC) for NARA have provided a useful opportunity for NARA to explore relevant technologies. However, the work has not informed many significant aspects of the ERA design, has not reduced the engineering risk of the program, and has not enhanced NARA's operational capabilities for running ERA systems.

The SDSC proof-of-concept demonstration projects have provided NARA with the opportunity to interact with the information technology (IT) community and to explore approaches for a production digital archiving system. The SDSC projects have demonstrated options for parts of a production digital archiving system, but NARA should not interpret these projects as solutions to digital archiving issues, as a substitute for gaining experience with operational pilots, or as a source of components of a production system. The areas where the SDSC work falls short of what would be accomplished through operational pilots include the following:

- *Scale.* The quantity of records tested in the SDSC demonstrations is small compared with the quantity of records that NARA anticipates ingesting.
- *Complexity*. The demonstrations addressed only a few relatively simple record types (such as e-mail and Senate legislative records), so the experience is not easily transferable to more complex problems.
- *Attention to trustworthiness issues.* For example, the SDSC work did not address requirements related to redundancy, integrity checks, or access controls.
- *Attention to operational matters.* The SDSC work should be understood as a demonstration of technology rather than as a prototype of an operational system or an operational pilot. The SDSC demonstrations shed little light on the operational issues, such as the work flow

associated with ingesting high volumes of varied records, that NARA will need to address in order to run ERA systems. Nor does this exploratory work substitute for having NARA staff work with a system in a production environment; users of SDSC's scientific data management systems, unlike the potential users of the ERA, are generally very savvy about technical matters and can turn to a highly proficient support staff.

Finally, the SDSC work emphasized a particular strategy for digital preservation: migrating records to XML-based formats. Although XML has important applications in archiving, the use of an XML document format does not solve the problem of format obsolescence, and it would be inappropriate to rely on a migration strategy alone for long-term preservation (see Recommendation 4 below). In addition, some aspects of the SDSC demonstrations were research work that attempted to express semantic constraints within records. This is a worthwhile long-term goal that may have some utility as a technique for ingesting or preserving certain types of records. However, this particular demonstration of "lifting knowledge" from a document is not persuasive; the technology is far from ready for inclusion in a production system.

ENGINEERING THE ERA (CHAPTERS 4 AND 5)

Finding 5. The broad principles and expectations that NARA has established thus far for the ERA are an insufficient basis for proceeding with its design, procurement, and operation.

NARA has thus far expressed the requirements and objectives for the ERA in very high-level terms. Some of these objectives stem directly from its statutory mandate, e.g., "preserve and provide access to any kind of electronic record."[4] It has also embraced the OAIS reference model to describe the high-level structure of the system.

More preliminary work is required in setting expectations for the system and estimating its size and scope before NARA can start procuring a workable production system. This includes (1) characterizing the electronic records that the ERA should be expected to ingest in the near term and (2) making pragmatic engineering decisions and defining realistic requirements and priorities. Only by jointly considering archival and technical concerns can NARA chart a course to meet its preservation mandate with achievable IT systems.

Recommendation 3. Before proceeding with design and procurement of the ERA, NARA should gather more data about the electronic records that it expects to preserve in the near future.

To formulate some of the quantitative and qualitative requirements of the ERA, NARA needs more information about the population of government records that it will hold and projections about how the records will be used. To guide the engineering of the ERA, and especially of its early versions, NARA should obtain or estimate these data now. None of these attributes will remain unchanged over the life of the system, but it is nonetheless important to develop the system based on these initial requirements.

When data are not available or are impractical to obtain, estimates should be prepared, justified, and made explicit; otherwise, a system design will reflect implicit estimates, which may be dangerously wrong. Importantly, the intent of this recommendation is not that the ERA program should be significantly delayed to conduct detailed surveys; only rough estimates are needed, and order-of-magnitude estimates will suffice in most cases.

Examples of the required data include these:

- Characterization of the population of digital records that will need to be preserved. How many government records, using what data types, will require preservation? How many records does NARA expect to receive in each future year? As remarked below, the ERA will need to prioritize handling of records based on archival and technical considerations. Both current data and forward estimates are required in order to make these decisions.
- Estimates of size and scaling trajectory. How much data will be stored in the ERA, and how will it grow over time? These estimates, which are needed to inform the technical structure of the system, may follow directly from estimates of the record population but may also be governed by the overall project plan, ingest rates, and other considerations.
- Mechanisms for delivering records to NARA. While today's and future records can be delivered to NARA using secure networking techniques, records generated over the past 30 years or less may reside outside network-connected systems on media that are rapidly becoming obsolete. How many records are stored on which media?
- Estimates of access rates. Since the ERA does not yet exist, access rates can only be estimated, perhaps based on experience with digital libraries (which might provide a better indication of user interest in online collections than would data on access to NARA's non-electronic records). The system will certainly need to be designed to increase access performance as demand increases, but even an initial system will require some estimate of access rates.
- Budget estimates. The quantitative estimates lead to estimates of costs for developing, procuring, and operating the ERA and are critical to making informed design decisions and investment trade-offs. If, for example, it turns out that the originally planned scope of the ERA would be unaffordable, criteria for preservation scheduling may have to be adjusted. Many design decisions concerning ingest processes and the amount of automation required for them will be driven by the cost of ingesting records of different types. The committee could not find much basis to support estimates for the cost of the ERA, nor could it determine whether current ERA plans are consistent with budgetary constraints.

Recommendation 4. NARA should address key design issues before commencing implementation and apply a pragmatic engineering approach to the ERA's development and evolution.

Although it may be tempting to speak in absolutes (e.g., "every important record will be preserved forever") when designing a system, engineering practice recognizes that there are objectives that are subject to constraints. Engineering the ERA will require specifying the objectives and constraints of such a system.

This report describes and the section below summarizes some of the important design issues that need to be addressed and provides some advice on how to think about them. In some cases, the committee's preliminary analysis led to design suggestions, but this is no substitute for comprehensive analysis by NARA. Key engineering tasks include these:

1. Prioritize the functions of the ERA and focus initial design on capabilities that permit rapid deployment of operational pilots. NARA's most basic requirement is to save bits for a hundred years or more. To achieve this requires a combination of careful technical and operational design based on extensive industry experience with robust storage systems of shorter life. Examples of measures to meet this goal are (1) redundant storage of bits at separate physical sites to survive physical destruction, (2) copying bits to new physical storage devices as existing devices age, (3) using storage systems with nonproprietary interfaces to prevent lock-in by any specific vendor, (4) careful system design and operation to guard against human error that might delete vital bits, and (5) diverse system implementations to guard against software errors that could lose data.

One way to gain experience is to build an effective bit storage capability that supports pilot programs to begin preserving records in the short term and that provides a critical foundation for future systems with broader capabilities.

Some basic ingest (i.e., intake of records and associated metadata) and access mechanisms are required in early ERA versions, but other functions that are less important can be deferred for later implementation, as long as the initial architecture, design concept, and implementation strategy are sufficiently flexible and evolvable and sufficient attention has been devoted to overall robustness, survivability, maintainability, and compatibility with critical long-term requirements.

2. Design for common cases. Given resource limitations, it is not reasonable to expect a system to preserve and provide equally good support for all records. A relatively small number of data types will likely support the majority of records that federal agencies are creating. It is also advisable that early system builds concentrate on a relatively small number of types. It will be necessary, therefore, to make some choices about what quality of service to provide for different types of record. To decide on the service level accorded each class of record, NARA will need to assess both technical and archival aspects of records. For some lower-priority formats, ERA support should, at least initially, be limited to capturing, storing, and providing access to the original bits and essential metadata.

3. Take pragmatic steps now to facilitate future access to records. NARA does not have to anticipate—or invest in—all the higher-level services that future users might want. Indeed, future archivists and researchers will be skilled in computing and thus will be more able to manipulate and interpret digital records. Also, many institutions, including NARA, share an interest in building an infrastructure of tools that support conversion, migration, and emulation; these will be available to NARA and its users.

The ERA should be designed so that certain fundamental information about a record is saved to enable future access. This pragmatic strategy includes the following elements:

- *Be neutral with respect to migration, emulation, or other approaches.* None of today's approaches—such as migration or emulation—for dealing with data type obsolescence have been perfected, nor has one emerged as accepted archival practice. Today, one should rely foremost on saving the original bits—even if one is unable to decode or render those bits when the records are ingested—together with additional information that facilitates their interpretation (preferred derived forms and essential metadata; see below). The viability of whatever strategies are used in the future will depend on the availability of the original bits.
- *Save records in "preferred derived forms" in addition to the original bits.* The derived forms simplify access to records because the formats are chosen pragmatically to be common, well-documented, and expected to last a long time. These derived forms can readily be created for many common record types by making use of existing export functions or conversion software. Derived forms are, however, no substitute for preserving and providing access to the original bits. (A related strategy that may be discussed in the committee's second report is to encourage agencies to create records in preferred formats at the outset.)
- *Do not rely primarily on a strategy of converting records to platform- and vendor-independent archiving formats in an attempt to avoid obsolescence.* This point follows from the previous two elements but is stated explicitly because it runs directly counter to the approach of converting all records to technology-independent formats, which is not likely to be effective. XML formats are often proposed for this role, but they cannot assume the role of the original data type because they cannot be relied on to faithfully encode all of the elements of all data types. By contrast, an XML derived form may be a very useful derived form that serves as an adjunct to saving the original data type.

- *Save essential metadata*. While it is often advisable to save as much metadata as possible, the most important metadata to save are those that cannot be derived from the record itself and thus would otherwise be lost (e.g., contextual metadata).
- *Save essential external references that are implicit or explicit in the record.* Digital files often refer to other digital objects, such as embedded images, tables generated by running some program, and files belonging to other organizations. The possibilities for cross-reference in digital files are far richer than in paper files, and rules will have to be developed to decide which cross-references should be preserved by copy, by reference, or both.
- *Archive as much information as possible about the software and work flow processes used to ingest the original records*. This information may be essential when future users of the archive wish to understand in detail how records have been processed. A desirable goal would be that the ingest process work flow be log-based and otherwise designed to facilitate analysis in case the preserved form of the record is later discovered to have been incorrectly ingested.

4. Safeguard the bits. The risks of the various possible causes of data loss—such as malicious acts, natural disasters, software bugs, human error, and hardware failures—should be assessed and used to make informed engineering cost-benefit trade-offs. A combination of appropriate system design and operational policies and procedures will be required. Measures to consider include redundancy (e.g., geographically distributed replicas), media refresh (copying data to new media before old media fail due to age), integrity checks (e.g., to verify the integrity of records received from agencies, to detect errors in data storage systems, and to protect against tampering), access controls (e.g., to control who can write or modify records and to protect classified or otherwise nonpublic records), and auditing. Some records may be deemed more important than others and will justify greater investment than others to ensure that they persist.

5. Select the appropriate storage media. The economics, performance, and robustness of all-disk storage systems have recently begun to exceed those of systems that include magnetic tape either as a primary storage medium or as a backup. While not yet common practice, it is likely that robust disk-only storage systems will become an attractive alternative to tape storage early in the life of the ERA. NARA should seriously consider such designs for the first ERA systems; they are much simpler than storage involving both tape and disk. Offline, possibly write-once, storage may continue to play some useful role in storing infrequently accessed records; the cost, complexity, performance, and reliability trade-offs associated with each technology option should be carefully considered. Even if the ERA does not initially eschew tape, it should be designed to make it easy to switch away from tape in the future.

6. Decide where to invest in access capabilities. Historically, the primary tool for finding physical (paper) records of interest has been the finding aid; finding aids, along with other surrogate records, are now used by computer systems that help people to find physical records. When the entire contents of digital records are available for computer processing, content-based retrieval techniques like full-text searching (which has proven to be a high-payoff, relatively low-cost method in other contexts) become possible. These will alter access interface designs, ingest, processing, access strategies, cost trade-offs, and even approaches to handling confidential or classified records.

7. Plan for consistent access to digital and physical records. Over time, it should become possible to use single, consistent access tools to search all the records in the custody of NARA, be they physical or digital. Indeed, over time, some current physical records may even be transferred to digital form. The ERA design needs to take this into account, and while such cross-collection capabilities would probably not be implemented in early iterations of the ERA, it is essential that the architecture recognize this long-term convergence.

INFORMATION TECHNOLOGY EXPERTISE (CHAPTER 6)

Finding 6. Greater information technology (IT) expertise is needed if NARA is to successfully design, acquire, and operate ERA systems.

Insufficient technical expertise at NARA is a major obstacle to successful development and acquisition of the ERA. Based on briefings and other interactions with NARA staff, the committee concludes that while there is recognition of the importance of the ERA program, few NARA staff members have experience with or fully understand the complexity of building and managing a program as challenging as the ERA. NARA today does not appear to have sufficient technical depth to assure success in launching the ERA program—that is, to define and manage the overall architecture, develop the appropriate request for proposals, evaluate technical responses, negotiate changes in the architecture with vendors, and manage the implementation of the system.

In addition to needing a quick ramp-up in the IT expertise necessary to oversee the early phases of procurement, NARA faces a longer-term need for a more pervasive culture change—IT skills related to preservation will need to be a core competence throughout the organization, on a par with its other institutional strengths. NARA recognizes the existence of this issue in its appointment of a change manager associated with the ERA program, but the difficulty in achieving this shift cannot be overestimated. It will not be possible to achieve the needed changes quickly; this pervasive change should be addressed in parallel with other facets of ERA development.

Recommendation 5. In order to pursue technical development of the ERA, NARA should first hire a small team of first-rate information technologists with systems design expertise.

The addition of a few employees with properly focused systems design expertise would greatly increase the likelihood that the ERA program will be successful.

Preparing the architectural design of the ERA requires first-class talent having both archival and IT expertise. Whether the architecture is defined by NARA staff (the preferred approach; see Recommendation 7) or contractors, the challenges of hiring qualified IT people are almost identical for these two approaches. If it is to be successful, the contracting approach requires NARA's expertise to equal that of the design contractors in order to determine whether a design will meet NARA's needs.

Contracting for system implementation once an architecture is defined likewise requires at a minimum an in-house contract monitoring staff (e.g., the contracting officer's technical representative) with IT expertise at least as good as that of the contractor's people. This expertise will be essential, in particular if NARA is to successfully pursue an iterative development approach (Recommendation 7).

Recommendation 6. To supplement its in-house expertise, NARA should recruit an advisory group of government, academic, and commercial experts with deep knowledge of digital preservation and IT system design.

A standing ERA advisory committee focused on digital preservation issues would provide an ongoing way to supplement NARA's IT capabilities. By drawing expertise from the range of digital preservation efforts under way in government, industry, and elsewhere, the advisory committee would allow NARA to learn from those efforts and to foster collaboration (e.g., on techniques, standards, or common components) were warranted.

STRATEGY FOR EVOLVING AND ACQUIRING THE ERA (CHAPTER 7)

Finding 7. Building the ERA as a conventional procurement is unlikely to succeed owing to the difficulty of accurately anticipating all system requirements. Instead, an iterative approach is needed.

Procurement of the ERA is fundamentally different from procurement of a payroll or other commonplace IT system. No one has built an ERA before, and its requirements are not yet completely understood. Also, some of the requirements—such as safeguarding bits with very high confidence—are stringent. As a result, the procurement should emphasize modularity, iteration, and working with vendors to define and evolve the system rather than arms-length specification and delivery of a completed, turnkey system.

The ERA will need to evolve, perhaps rapidly, during its early years as technical and operational requirements are modified by experience. Later in its life, the ERA may evolve more slowly as new needs are identified and as old hardware and software components are replaced or upgraded.

Recommendation 7. The ERA should be designed as a modular system that can be built, maintained, modified, and evolved incrementally, subject to an overall architecture.

A proper modular design allows components to be upgraded without disrupting the operation of the system. An overall structure for the ERA is suggested by the OAIS reference model, but a design for the ERA will need to be much more detailed in order to exploit the benefits of modularity. A modular structure would make it easier to use COTS components for the ERA.

The system's architecture ensures that the pieces fit together, that the system can be incrementally modified, and that it can evolve over time. Interfaces between major parts should be specified using an open approach that allows multiple vendors to supply components over a long lifetime. The architecture is itself subject to evolution over time as requirements are better understood or new requirements emerge. However, devising a good initial architecture is very important as the program's success will be sensitive to the nature of the architectural decisions that are made early on. It is, for example, far harder to evolve the modular structure than to evolve individual modules.

The architecture of the ERA system(s) should be "owned" (specified and evolved over time) by NARA. This would help ensure that NARA understands the implications of alternative proposals, reduces its dependence on vendors and the risk of proprietary lock-in, and understands the limitations and strengths of systems that vendors deliver. Preparing the architectural design of the ERA requires first-class talent having both archival and IT expertise. So too does evolving it over time to meet new requirements.

A far poorer alternative to a NARA-owned architecture is for NARA to contract for one or more architectural designs. In this case, it may be worthwhile to obtain several proposals, because different contractors may have different opportunities or ideas for incorporating COTS elements. This alternative also requires first-class technical talent in order to specify the scope of a design contract, to evaluate resulting designs, and to proceed with acquiring and evolving a system.

Recommendation 8. NARA should begin development of the ERA with a small number of focused pilot production systems designed to gain early experience and to converge ultimately into a smaller number of more comprehensive systems.

NARA should concurrently develop and deploy small, focused systems that rapidly build operational experience. All of these systems should be built within a common architectural framework (Recommendation 7), so that they may eventually coalesce into a smaller number of more comprehensive systems as experience and confidence grow. It is especially important that the data model—the data types and related metadata—conform to the architecture so that the digital data obtained by ingesting records into one of the early systems will carry forward into future evolutions.

The initial systems should be selected and scoped for rapid deployment—this is the key to gaining early experience to inform the requirements of later systems. The following are some examples of limited-scope systems that might be considered for early pilots:

- *U.S. State Department diplomatic cables.* NARA is preparing to acquire a collection of diplomatic cables, which are simple structured text files, in digital form. Ingest might include automatic extraction of metadata from the cables; access might include full-text search or other methods appropriate to the collection. For quickest deployment, NARA might consider making these records available using software already developed for operating a digital library.
- *Records at the National Personnel Records Center.* There is interest in preserving large but homogeneous collections of official military records scanned in TIFF image format when they are transfered to NARA's National Personnel Records Center. Confidentiality considerations and the imperative to provide ready access to veterans or next-of-kin would require careful attention to access controls.
- *E-mail from the Clinton administration held by the Clinton Presidential Center.* Metadata could be extracted from the e-mail headers, full-text search could be provided, and so on. The presence of attachments would permit gaining experience with preserving and providing access to a broad range of relatively contemporary data types.

These three examples illustrate collections that could be organized and made available quickly. Although these collections might lack the scale of the eventual ERA, early deployment of systems to preserve and access them would yield important operational experience for NARA and avoid costly mistakes in later, more complex systems.

Experience with early systems can be expected to lead to changes to the ERA architecture and to the substantial refinement of requirements for subsequent, more comprehensive systems. Managing the initial architecture, the first system deployments, the learning from early operations, and the revisions to architecture and specifications, and evolving the ERA will be the task of NARA's augmented IT staff.

NOTES

1. National Archives and Records Administration (NARA). 2000. *Ready Access to Essential Evidence: The Strategic Plan of the United States National Archives and Records Administration 1997-2002* (revised 2002). National Archives and Records Administration. Government Printing Office, Washington, D.C.
2. "Data types" is a more general term than "file formats," though the latter may be more familiar.
3. Consultative Committee for Space Data Systems (CCSDS). 2002. *Reference Model for an Open Archival Information System (OAIS)*, CCSDS 650.0-B-1 (Blue Book). CCSDS Secretariat, National Aeronautics and Space Administration, Washington, D.C. January. Available online at <http://wwwclassic.ccsds.org/documents/pdf/CCSDS-650.0-B-1.pdf>.
4. The basic goals of NARA and the ERA program, as expressed in NARA documents, are presented in Appendix A [not reprinted here].

Appendix C

October 16, 2003, Letter Report to the National Archives and Records Administration

October 16, 2003

Dr. Kenneth Thibodeau
Director, Electronic Records Archives (ERA) Program Management Office
The National Archives and Records Administration
8601 Adelphi Road
College Park, MD 20740

Dear Dr. Thibodeau:

In this letter,[1] the National Research Council's Committee on Digital Archiving and the National Archives and Records Administration, chartered to study the Electronic Records Archives (ERA) program, offers comments on the ERA program's draft request for proposals (DRFP)[2] and the attached requirements document (RD)[3] and ERA deployment concept (ERA DC)[4] that were made available for public review in August 2003.

The comments below elaborate on issues discussed in the committee's first report, *Building an Electronic Records Archive at the National Archives and Records Administration: Recommendations for Initial Development* (hereafter referred to as the committee's first report),[5] tying the issues specifically to the DRFP, which was not available in early 2003 when the committee completed its report. A meeting with the ERA team on August 14, 2003, helped the committee to better understand the philosophy underlying the DRFP and also informed the comments below.

The committee supports the National Archives and Records Administration's (NARA's) goal of building new electronic records preservation capabilities quickly to meet NARA's and the nation's needs to preserve federal records (committee's first report, Findings 1 and 2, pp. 1-2). The committee understands that the general objective of the DRFP is to spell out these needs and to seek proposals that offer thoughtful and innovative ways to meet them (thus challenging industry), rather than to prescribe too narrowly how the ERA must be designed.

As detailed in item 1 below, the committee is especially concerned that the software development process outlined in the DRFP and RD does not reflect Finding 7 of the committee's first report (p. 11), "Building the ERA as a conventional procurement is unlikely to succeed owing to the difficulty of accurately anticipating all systems requirements. Instead, an iterative approach is needed."

The committee makes the following specific suggestions for how the DRPF might be improved:

1. **Adopt an iterative design approach.** The DRFP is written assuming a classic waterfall design methodology in which requirements are established early and development proceeds in a serial fashion. The committee's first report advocated a quite different approach—spiral design—that makes use of early prototyping and continuous refinement of requirements (Finding 7, p. 11; Recommendation 8, p. 12; and discussion, pp. 67-70). The DRFP asks vendors to provide information on their proposed use of prototyping,[6] but prototyping does not receive as much emphasis as it should.

Discussion with NARA indicated that the two design phases and five deployment increments described in the DRFP are to be treated somewhat like iterative design: at each stage, a new set of requirements will be prepared, and parts of the system may be revised and new parts implemented for the first time. Discussion also indicated that easier, less risky system parts would be designed first and the more difficult ones delayed in the expectation that new and better approaches would emerge. This strategy will not reap the full benefits of iterative design, which requires learning early on from prototypes and an ability to evolve requirements and improve subsequent designs as more is learned.

For example, a number of important technical details—e.g., file formats supported, essential metadata tags selected, and the data model used to save ingested records—are likely to require iterative development. Some of these design decisions (e.g., the data model) will be much easier to change when the system is young and has not yet ingested a large amount of data. In fact, some revisions may require reading and reformatting the contents of the entire archive.

Also, the mechanisms for ingesting new records are likely to require extensive iterative design. The degree of automation, human interface, and resulting productivity of the ingest process will be critical to the success of the system (committee's first report, p. 49). Designing a good user interface is hard even when the needs and skills of users are well understood; neither is the case here. The DRFP is silent on the qualifications of individuals required to operate and maintain the ERA; such aspects as the human error rate will be critical to the long-term viability and cost of the system.

2. **Specify that an explicit threat model be developed early in the ERA's life cycle.** The DRFP makes occasional mention of measures that might help in averting threats (physical distribution of storage sites, user authentication, audit logs, etc.), but it includes no overall requirement that the system be capable of surviving an attack or accident.[7] Records stored in the ERA will be kept for hundreds of years and must be protected against loss due to natural disasters, hardware and software failures, operator errors, and potential attacks from both inside and outside NARA (committee's first report, Recommendation 4, p. 9). The committee's first report advocated that NARA develop an explicit threat model and evaluate ERA designs against the model (p. 54). Retrofitting of security measures has repeatedly been demonstrated to fail. Because security should be designed in from the beginning (committee's first report, p. 53), a threat analysis should be done as early as possible in the ERA life cycle. Moreover, contractors may propose quite different system designs depending on the threat model that is developed.

3. **Tighten up requirements related to security, integrity, survivability, reliability, and robustness.** In addition to specifying that a threat analysis be done at the outset, the RD should be revised to clarify a number of additional points:

- *ERA 1.2*[8] should include a requirement that record transfers can be integrity-checked (committee's first report, Recommendation 4, p. 9 and p. 56). ERA 1.2 mentions authorization and ERA 13.7 mentions authentication, but these safeguards are distinct from integrity-checking.

- *Protection should be provided against human error and malicious insiders.* Any request that would delete or permanently alter data in the archive should require confirmation by two or more separate humans who are authorized to act on the request, a requirement that is not in the current RD.
- *ERA 13.4, "The system shall support virus detection," and ERA 13.5, "The system shall support virus elimination."* How will the system handle a record that appears to include a virus? Is the virus part of the record? Whose responsibility is it to deal with embedded viruses—NARA's or the end-user's?
- *ERA 13.8, "The system shall provide for backup of ERA."* How does this requirement relate to the deployment strategy of "safe stores"? Is such backup in addition to safe stores, or do safe stores meet this requirement?
- *ERA 14.5, "The system shall use templates to check the authenticity of transferred electronic records."* This requirement is unclear, given the use of the vague term "template." What information contained in a template can be used to check authenticity? How? Also, it might be better to require simply that the authenticity of records be checked and leave it to contractors to develop a suitable approach.
- *ERA 18.10.2, "The system shall output certified copies of electronic records in formats selectable by the user from available choices."* What is a "certified copy"? Will these copies be electronic files "signed" by NARA? Is the contractor being asked how to certify copies and how to record such certification?
- *ERA 21.1, "The system shall register users."* This requirement appears to imply that access cannot be anonymous. Is this necessary or desirable?
- *RD, section 2.7.4 second paragraph.* The term "non-repudiation" is used incorrectly here (either that or the committee misunderstands the paragraph).

4. **Explicitly require a levels-of-service approach.** Service-level differentiation is a practical necessity (committee's first report, p. 26). If the ERA must be able to accept all data types supplied by agencies, then "ideal" preservation solutions cannot practically be developed for the full spectrum of data types expected.

The committee has observed a growing recognition at NARA that different record types or contexts will require different levels of service. The DRFP includes references to this approach,[9] but the full ramifications of service-level differentiation are not yet clearly expressed, and the level-of-service approach does not appear in the requirements themselves. In the spirit of seeking industry ideas, the DRFP should include an express requirement to classify records as to levels of service, so that proposals will include mechanisms to help set, enforce, and revise the service levels associated with records or groups of records. Just as templates may be required to streamline the ingest of common record types, so also a codification of service levels may be an important part of the ERA's structure.

ERA 18.5.2, "The system shall provide the capability to electronically present electronic assets independently of the software with which they were created," should also be clarified and reconciled with the notion of providing levels of service. Presumably the intent of ERA 18.5.2 is that lack of access to the software that created a record should not preclude access to the record itself. Yet if the level of service for a particular record provides for keeping only the bits—because, for example, the data type is so seldom used or so unimportant that no attempt has been made to emulate or migrate the record or do any other form of conversion—then "present" must mean "deliver the bits" rather than "display an image."

Finally, Table C-4, "ERA Supported Data Types,"[10] should be filled in more completely to reflect (1) the data types that NARA has already announced it will support, such as Portable Document Format (PDF), (2) a concrete approach to requirement ERA 28, "The system shall accept all types of electronic records," and (3) varying levels of service and how they would be

provided for different record types. The ERA must be designed, at the outset, to support a broad range of record types, even if implementations of some types are deferred.

5. **Avoid overspecification.** In some places, the DRFP calls for specific approaches to implementation rather than asking for proposals for an overall system that exhibits particular desired behaviors. Calling for the latter would be consistent with the strategy presented at the August 14, 2003, meeting, which is to challenge industry to propose innovative solutions. Specific instances of the requirements specifying particular implementation approaches rather than input-output behaviors include the following:

- *Multiple references to "persistent data formats."*[11] It would be best to omit these references, because whether and how to employ persistent data formats is an internal implementation decision. Instead, NARA can evaluate proposed designs on the basis of how they handle the obsolescence of data types, the degree to which they introduce errors, and so on. Although the use of persistent formats may ultimately be the best strategy for at least some types of records, it is not the only possible approach nor necessarily the best in all circumstances.[12] The requirements should be neutral with respect to how a contractor may propose designing a system that will retain the requisite level of functionality and access over time.
- *ERA 11.5, "The system shall store electronic records such that an individual electronic record does not span media volumes."* This decision should be up to the implementers as long as the input-output requirements are met. In particular, this requirement would seem to rule out redundant array of independent disks (RAID) technology and related techniques for file-system redundancy, which are arguably applicable to the ERA.
- *ERA 9.2.7.2, "The system shall maintain an archive file directory defining the physical locations of all electronic records within the system."* This requirement also appears to overspecify the design. A requirement to be able to obtain the physical location of electronic records in the ERA should be made into an external (output) requirement.
- *"ERA will support automated media maintenance and tools to recover data from failed media."*[13] Does this requirement apply to ingesting records from media provided by others, or to media used within the ERA to store records? (See also ERA 11.4.) If the latter, this requirement would intrude on a contractor's ability to provide the best technologies for storing data for long periods—chiefly, using redundancy techniques in such a way that recovering data from failed media is unnecessary. If the phrase "automated media maintenance" refers to robotic tape libraries, it, too, overconstrains the solution.
- *Detailed requirements on search characteristics included under ERA 17.5.* Is this detailed specification necessary? Proposals should include state-of-the-art searching capabilities. (See also points 7 and 8, below.)

6. **Strengthen the requirement for the ERA to preserve the original bit stream.** An essential input-output behavior is that the ERA be able to ingest, preserve, and produce the original bit stream of a record in its native format (committee's first report, p. 31). Although the August 2003 discussion indicated that this is the intent, the committee believes this requirement may not be stated sufficiently strongly in the RD. Perhaps requirement ERA 7.2, "The system shall support retaining data files in the formats in which they were ingested," is intended to address this requirement. If so, it should be strengthened to specify that the original bit stream of each record must be saved (the requirement as currently worded says only that the original format be used, not that the bits must remain unaltered). (Of course, depending on the level of service deemed appropriate for a particular record or class of records, additional preservation measures might well be undertaken as specified in the record's preservation plan.)

7. **Require more flexible search facilities.** The RD as written assumes that there is a single search mechanism for the entire archive. But different collections or record types might be more accessible with quite different search engines or techniques. The committee recommends that the ERA be structured so that several search techniques can be accommodated, and so that search software can be easily changed or replaced over time. (Search technologies are discussed briefly in the committee's first report, pp. 51-53.)

8. **Explicitly require a federation interface that supports search and retrieval across multiple archives.** Researchers often search more than a single archive and will benefit from federated search and access mechanisms whereby several archives can be accessed at once. (This capability to search for and access items across institutions would be in addition to whatever federation techniques are used within the ERA to manage storage.) If ERA, as the nation's archive of electronic records, offers a federation mechanism and invites others to join, it will induce others to do so as well. The RD should challenge industry to design a federation interface based on approaches being developed or prototyped by others (e.g., the digital library community).

9. **Plan for increasingly distributed operations.** While the core of the ERA will be distributed among several sites and closely managed by NARA, some processes may be deployed elsewhere. For example, it would seem essential to permit record ingest to be performed within a client agency, or by a contractor working for the agency and/or NARA. Over time, agencies may come to play a greater role in record ingest, and NARA's role will be to provide the ingest software or specifications, and to provide careful automatic checking of the records that have been ingested, together with associated data (data currently entered in Standard Form 258, metadata tags, file format integrity, authenticity checks, etc.). Another good example of the value of distributed operations is the opportunity to leverage externally supplied preservation tools, either for file migration or for on-the-fly presentation of an obsolete format.

The ERA should, therefore, be designed from the outset so that certain parts can be operated at geographically separate sites to accommodate flexible workflows. Perhaps ERA 1.8 and ERA 1.9 are intended to address distribution, but it is not clear. ERA should, for example, be designed so that NARA employees or contractors physically outside a NARA facility can have secure access to the ERA. This capability will enable contracting out various processing tasks and also allow NARA employees to work from home. (Note that a remote access requirement also affects ERA 20; the user interface should permit operation from, say, standard PCs on the Internet.)

Any need for other agencies or contractors to be able to examine records for sensitive content (ERA 12.4) will have an impact on workflow and other matters. Is there a need to record who (or which automated process) made a decision regarding a record's sensitivity?

Of course, a careful threat analysis (see point 2, above) with respect to distributed operations is essential, and security requirements associated with distributed operations should be included under ERA 13.

10. **Specify more precisely the acceptable update lag between primary and "safe store" data repositories.** The ERA DC correctly notes that there will be a lag while data stores synchronize.[14] For system design to proceed, this lag will have to be specified more precisely: if the lag is on the order of seconds, large instantaneous bandwidth may be required to connect the safe store sites in order to communicate updates, whereas if the lag can be longer, then the communications need only meet average ingest rates and might even use mailed media rather than network connections for synchronization.

11. **Provide better definitions for key terms.** Despite an impressive glossary in RD Appendix B, some important terms are used without clear definition in the DRFP. The DRFP should introduce better definitions or cite literature in which these terms are explored further. For example:

- *Essential characteristics*—A clear idea of what this phrase means in practice would be valuable. For example, will the need for strong functional equivalence arise? Some concrete examples might be helpful.
- *Persistent format*—This term has been introduced in several papers written in conjunction with the ERA program, but the committee found especially helpful the clarification offered at the August 2003 meeting—that is, the use of persistent formats as a form of migration that tries to minimize the number of transformations.
- *Self-describing format*—This is a slippery term: every description is written in a "language" that itself must be described, thus begetting an infinite recursion. Again, examples would help to clarify a non-obvious term. See also ERA 11.3.
- *Conceptual search*—This phrase means different things to different people. If the intent is to distinguish between controlled-vocabulary (metadata) search and full-text search, it would be sufficient to say simply "full-text search." If something more elaborate is meant, the requirement is asking for a capability beyond the state of the art.
- *Template*—The term is defined in operational terms, but not clearly with respect to its contents or (detailed) role.
- *Electronically*—This term is used in an unnatural way in several places in the RD. For example, ERA 14.2 says, "The system shall provide the capability to accept transfers electronically." Why not simply say "via a network connection"? Or is the term "electronically" meant to include disk and tape, too? Similarly, ERA 18.1 says, "The system shall be capable of electronically presenting all electronic record types." Should the wording be "delivering via a network connection" instead of "electronically presenting"? For some records, the level of service may be limited to delivering the original bit stream. In each case, a more explicit definition would help.

Sincerely,

Robert F. Sproull, *Chair*
Committee on Digital Archiving and the National Archives and Records Administration

NOTES

1. Support for this project was provided by the National Archives and Records Administration under Contract No. NAMA-02-C-0012. Any opinions, findings, conclusions, or recommendations expressed in this letter are those of the authors and do not necessarily reflect the views of the organization that provided support for the project.

This letter report has been reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise, in accordance with procedures approved by the National Research Council's Report Review Committee. The purpose of this independent review is to provide candid and critical comments that will assist the institution in making its published report as sound as possible and to ensure that the report meets institutional standards for objectivity, evidence, and responsiveness to the study charge. The review comments and draft manuscript remain confidential to protect the integrity of the deliberative process. We wish to thank the following individuals for their review of the report: William Y. Arms, Cornell University; David. D. Clark, Massachusetts Institute of Technology; Judith L. Klavans, Columbia University; MacKenzie Smith, MIT Libraries; and J. Timothy Sprehe, Sprehe Information Management Associates. Although these reviewers provided many constructive comments, they were not asked to endorse the conclusions or recommendations, nor did they see the final draft of the report before its release.

The review of this report was overseen by Robert J. Spinrad, Xerox Corporation (retired). Appointed by the National Research Council, he was responsible for making certain that an independent examination of this report was carried out in accordance with institutional procedures and that all review comments were carefully considered. Responsibility for the final content of this report rests entirely with the authoring committee and the institution.

2. Electronic Records Archives Program Management Office, National Archives and Records Administration (NARA). 2003. *Electronic Records Archives Draft Request for Proposal (DRFP)*. NARA, College Park, Md., August 5. Available online at <http://www.archives.gov/electronic_records_archives/acquisition/draft_rfp.html>.

3. Electronic Records Archives Program Management Office, National Archives and Records Administration (NARA). 2003. *Electronic Records Archives Requirements Document (RD)*, (DRFP, Section J, Attachment 2). NARA, College Park, Md., July 31. Available online at <http://www.archives.gov/electronic_records_archives/acquisition/draft_rfp.html>.

4. Electronic Records Archives Program Management Office, National Archives and Records Administration (NARA). 2003. *ERA Deployment Concept (DC)*, (DRFP, Section J, Attachment 17). NARA, College Park, Md., July 29. Available online at <http://www.archives.gov/electronic_records_archives/acquisition/draft_rfp.html>.

5. National Research Council. 2003. *Building an Electronic Records Archive at the National Archives and Records Administration: Recommendations for Initial Development*, Robert F. Sproull and Jon Eisenberg, editors. The National Academies Press, Washington, D.C. The committee's final report will be issued in early 2004.

6. DRFP, subsection 3.1.2, p. L-18.

7. Storage robustness issues are discussed in the committee's first report at pp. 42-44, security and access controls are discussed at pp. 53-55, and record integrity is discussed at pp. 56-57.

8. References in the form (ERA x.y) are to numbered requirements appearing in the RD, pp. 19-48.

9. "ERA must be capable of providing different levels of service," RD, p. 8. "Levels of service—the ability of ERA to provide different capabilities for different records," RD, p. B-6.

10. RD, p. C-4.

11. "For permanent records—those preserved forever—and for some temporary records which need to be kept for lengths of time that exceed several generations of information technology, it will be necessary to transform the records from the formats in which they were received to persistent formats," RD, p. 10. "NARA's goal is to preserve electronic records in persistent formats that will enable access to authentic electronic records indefinitely into the future," RD, p. 14.

12. The committee's first report cautions against relying primarily "on a strategy of converting records to platform- and vendor-independent archiving format to avoid obsolescence" (p. 32) and discusses various approaches to preservation (p. 8 and pp. 32-34).

13. RD, p. 15, last paragraph under heading 2.7.3.

14. DC Section 4.0, "Synchronization with the operational record stores," p. 8.

D

Committee Member and Staff Biographies

COMMITTEE

Robert F. Sproull, *Chair*, vice president and fellow at Sun Microsystems, founded and led the Massachusetts branch of Sun Microsystems Laboratories for more than 10 years. Now he serves in a research role. Since undergraduate days, he has been building hardware and software for computer graphics: clipping hardware, an early, device-independent graphics package; page-description languages; laser printing software; and window systems. He has also been involved in very large scale integrated (VLSI) circuit design, especially of asynchronous circuits and systems. Prior to joining Sun, he was a principal with Sutherland, Sproull and Associates; an associate professor at Carnegie Mellon University; and a member of the Xerox Palo Alto Research Center. He is a coauthor with William Newman of the early text *Principles of Interactive Computer Graphics*. He is a member of the National Academy of Engineering and has served on several National Research Council (NRC) committees, including the CSTB committee that produced *Making IT Better*.

Howard Besser is an associate professor in the School of Education and Information Studies at the University of California, Los Angeles, where he teaches courses and does research on digital longevity, multimedia, image databases, digital libraries, intellectual property, instructional technology, and the social and cultural impact of new information technologies. For the past decade, Dr. Besser has been involved in digital longevity issues for cultural heritage materials. He served on the 1995 Commission on Preservation and Access Task Force on Preservation of Digital Information, and wrote the background paper on digital longevity for the Getty Research Institute's 1998 Time and Bits Digital Longevity Conference. In 2001 he served on the University of California's Task Force on Digital Preservation and Archiving, and he ran a series of workshops on digital longevity for the San Francisco Museum of Modern

Art. In the past few years, Dr. Besser has authored three pieces on digital longevity. He recently joined the InterPARES Project (International Research on Permanent Authentic Records in Electronic Systems) and will be the founding director of New York University's new master's degree in moving image archiving. Dr. Besser served on a previous NRC/CSTB committee that examined intellectual property in the digital age and produced the report *The Digital Dilemma*. He received his B.A. (1976), M.L.S. (1977), and Ph.D. (1988) from the University of California, Berkeley.

Jamie Callan is an associate professor at Carnegie Mellon University. His background is in information retrieval and machine learning. His recent research on information retrieval addresses automatic database selection, high-speed adaptive information filtering, novelty detection, question answering, and information literacy in K-12 education. His earlier research studied architectures for large-scale information retrieval and filtering systems, first-generation Web-search systems, methods of improving search accuracy, and integration of text search with relational database systems.

Charles Dollar is a consultant in archiving and records management, currently addressing the archival preservation of Smithsonian Institution Web resources. He has 20 years' experience working at the National Archives and Records Administration (NARA). While he was on the NARA staff, Dr. Dollar organized and directed the first electronic records program for the federal government and subsequently had a major role in research projects on digital technology standards, digital storage media, and digital imaging applications in electronic archiving. In 1994 Dr. Dollar joined the graduate faculty of the School of Library, Information, and Archival Studies at the University of British Columbia, where he taught in the Archival Studies Program. In 1999 he joined Cohasset Associates as a senior consultant. Dr. Dollar's publications, reports, and consultant studies are noted for their clarity in explaining digital technology issues and for providing practical guidance on such matters as planning and implementing effective life-cycle electronic records management programs with an emphasis on electronic archiving. Dr. Dollar is a fellow of the Society of American Archivists. He has a Ph.D. in history from the University of Kentucky.

Stuart Haber is a researcher at Hewlett-Packard Laboratories. Along with Scott Stornetta, he was a co-founder of Surety, Inc., which was spun off by Bellcore in 1993 to commercialize the secure digital time-stamping technology that the two of them developed as researchers at Bellcore (now Telcordia Technologies). In addition to Bellcore and Surety, he has also worked as a researcher in STAR Laboratory, the research arm of InterTrust Technologies, which provides digital rights management systems. Dr. Haber received his B.A. from Harvard University and his M.S. from Stanford University, both in mathematics, and his Ph.D. in computer science from Columbia University in 1988. He has lectured and published scientific articles on several practical and theoretical aspects of cryptology, on the theory of computing, and in electrical engineering.

Margaret Hedstrom is an associate professor in the School of Information at the University of Michigan, where she teaches in the areas of archives, electronic records management, and digital preservation. Prior to joining the faculty at Michigan in 1995, she worked for 10 years at the New York State Archives and Records Administration, where she was chief of State Records

Advisory Services and director of the Center for Electronic Records. Dr. Hedstrom earned her master's degrees in library science and history and a Ph.D. in history from the University of Wisconsin-Madison. She wrote a dissertation on the history of office automation in the 1950s and 1960s. She is a fellow of the Society of American Archivists and was the first recipient of the annual Award for Excellence in New York State Government Information Services. She served on the CSTB study committee that wrote *LC21: A Digital Strategy for the Library of Congress*. Dr. Hedstrom is widely published on various aspects of archival management, electronic records, and preservation in digital environments, and she has served as a consultant to many government archival programs. Her current research interests include digital preservation strategies, the impact of electronic communications on organizational memory and documentation, and remote access to archival materials.

Mark Kornbluh is an associate professor of history at Michigan State University and the director of MATRIX: The Center for Humane Arts, Letters, and Social Sciences On-line. Dr. Kornbluh also serves as executive director for H-Net: Humanities and Social Sciences OnLine, an international scholarly society composed of more than 140 online networks, edited by scholars in North America, Europe, Africa, and the Pacific. He is the principal investigator on research and education projects funded by the National Endowment for the Humanities, the National Science Foundation, the U.S. Department of State, the U.S. Agency for International Development, and the Andrew Mellon and Ford Foundations. The impact of new communications technologies on scholarly research and education is the prime focus of Dr. Kornbluh's research interests. In addition to awards for excellence in teaching, he has been the recipient of a number of grants to create multimedia teaching tools and resources. He has also authored a book on voter participation in the early part of the 20th century, *Why America Stopped Voting: The Decline of Participatory Democracy and the Emergence of Modern Electoral Politics, 1880-1918* (New York University Press). He earned a B.A. in history and political science from the University of California, Berkeley (1977) and an M.A. (1979) and a Ph.D. (1988) in history from the Johns Hopkins University.

Raymond Lorie is a research staff member at the IBM Almaden Research Center in San Jose, California. He graduated as Ingénieur Civil Electricien-Mécanicien from the University of Brussels, Belgium, and joined IBM Belgium in 1960. On assignment in Cambridge, Massachusetts he developed one of the very first access methods for relational data. He joined IBM Research in San Jose in 1973 and did pioneering work in relational database systems. Dr. Lorie was a major contributor to the architecture and implementation of System R, the early research prototype that became the precursor of the relational products. In later years, he managed projects on the application of the relational technology to engineering and other nonbusiness areas and on the use of parallelism in database systems. At various points in his career, Dr. Lorie dealt with document processing, from the design of a typesetting system in the early days to the co-invention of the Geography Markup Language (GML), the application of contextual knowledge to improve automatic recognition of characters, and finally the development of a technology to preserve digital documents for the very long term. He is a fellow of the Association for Computing Machinery (ACM). For several years, he taught a graduate class on database topics at the University of California, Santa Cruz. His work on System R brought him an IBM Corporate Award; he also shared with colleagues from IBM and the University of California, Berkeley, the 1988 ACM System Award for developing the relational technology.

Clifford Lynch has been the director of the Coalition for Networked Information (CNI) since July 1997. CNI, jointly sponsored by the Association of Research Libraries and Educause, includes about 200 member organizations concerned with the use of information technology and networked information to enhance scholarship and intellectual productivity. Prior to joining CNI, Dr. Lynch spent 18 years at the University of California Office of the President, the last 10 as Director of Library Automation. He is an adjunct professor at Berkeley's School of Information Management and Systems. He is a past president of the American Society for Information Science and a fellow of the American Association for the Advancement of Science (AAAS) and the National Information Standards Organization. Dr. Lynch currently serves on the Internet 2 Applications Council and the National Digital Preservation Strategy Advisory Board of the Library of Congress. He was a member of the National Research Council/CSTB committee that published *The Digital Dilemma: Intellectual Property in the Information Age*, and he served on CSTB's Committee on Broadband Last Mile Technology, which produced the report *Broadband: Bringing Home the Bits*. He holds a bachelor of arts in mathematics and computer science from Columbia College, a master of science in computer science from the Columbia University School of Engineering, and a Ph.D. in computer science from the University of California, Berkeley.

Jerome H. Saltzer is a professor of computer science, emeritus, at the Massachusetts Institute of Technology (MIT). Since 1966, he has taught at MIT, where he helped formulate the undergraduate curriculum in computer science and developed the core subject on the engineering of computer systems. At the MIT Laboratory for Computer Science he developed RUNOFF, the ancestor of most typesetting formatters. It, together with the context editor TYPSET, constituted one of the first widely used word-processing systems. He participated in the refinement of the Compatible Time-Sharing System (CTSS) and was involved in all aspects of the design and implementation of the Multiplexed Information and Computing Service (Multics), including the design of the first kernel thread package, the first time-of-century clock and, in the early 1970s, a project to develop what would today be known as a microkernel. Together with David Clark and David Reed, Professor Saltzer articulated the end-to-end argument, a key organizing principle of the Internet. More recently, his research activities have involved the design of a token-passing ring local area network, networking of personal computers, and designing the electronic library of the future. From 1984 through 1988 he was technical director of MIT's Project Athena, a system for undergraduate education comprising networked engineering workstations, and probably the first successful implementation of the network computer. Throughout this work, he has had a particular interest in the impact of computer systems on society, especially on privacy and the risks of depending on fragile technology. In September 1995, Professor Saltzer retired from the full-time faculty. He continues to write and teach about computer systems part-time from his MIT office. Professor Saltzer is a member of the National Academy of Engineering and a former member of the Computer Science and Telecommunications Board. He served on CSTB's Committee on Information Technology Strategy for the Library of Congress. He is a fellow of the Institute of Electrical and Electronic Engineers (IEEE) and the AAAS, a member of the Association for Computing Machinery, Sigma Xi, Eta Kappa Nu, and Tau Beta Pi; he is also a member of the Mayor's Telecommunications Advisory Board for the City of Newton, Massachusetts. He received the degrees of S.B. in 1961, S.M. in 1963, and Sc.D. in 1966, from MIT, all in electrical engineering.

Margo Seltzer is the Herchel Smith Professor of Computer Science and the associate dean for computer science and engineering in the Division of Engineering and Applied Sciences at Harvard University. She is also the founder, chairman of the board, and chief technical officer for Sleepycat Software, which produces Berkeley DB, an embedded, transactional data store. Her research interests include file systems, databases, and transaction processing systems. She is the author of several widely used software packages, including database and transaction libraries and the 4.4BSD log-structured file system. Dr. Seltzer spent several years working at start-up companies designing and implementing file systems and transaction processing software and designing microprocessors. She is a Sloan Foundation Fellow in Computer Science, a Bunting Fellow, and was the recipient of the 1996 Radcliffe Junior Faculty Fellowship, and the University of California Microelectronics Scholarship. She is recognized as an outstanding teacher, winning the Phi Beta Kappa teaching award in 1996 and the Abrahmson Teaching Award in 1999. Dr. Seltzer received an A.B. degree in applied mathematics from Harvard/Radcliffe College in 1983 and a Ph.D. in computer science from the University of California, Berkeley, in 1992.

Robert Wilensky is professor in the Computer Science Division and the School of Information Management and Systems at the University of California, Berkeley, where he has spent his entire professorial career. While at Berkeley, Dr. Wilensky has served as director of the University of California, Berkeley/Hewlett Packard Science Center, director of the Cognitive Science Program, director of the Artificial Intelligence Research Project, and as chair of the Computer Science Division from 1993 to 1997. Dr. Wilensky has published numerous articles and books in the area of artificial intelligence, planning, knowledge representation, natural language processing, and digital information systems. He is currently principal investigator of the University of California, Berkeley's Digital Library Project. Dr. Wilensky is a fellow of the American Association for Artificial Intelligence and an ACM fellow.

STAFF

Jon Eisenberg is a senior program officer with the Computer Science and Telecommunications Board of the National Research Council. At CSTB, he has been study director for a diverse body of work, including a series of studies exploring networking technologies and Internet and broadband policy. Current studies include an examination of emerging wireless technologies and spectrum policy and a review of the National Archives and Records Administration's digital materials preservation strategy. From 1995 to 1997 he was a AAAS Science, Engineering, and Diplomacy Fellow at the U.S. Agency for International Development where he worked on environmental management, technology transfer, and telecommunications policy issues. He received his B.S. in physics with honors from the University of Massachusetts at Amherst in 1988 and his Ph.D. in physics from the University of Washington in 1996.

Jennifer M. Bishop, program associate, has been with the Computer Science and Telecommunications Board of the National Research Council since 2001. She provides research assistance for several studies, including Telecommunications Research and Development and Digital Archiving and the National Archives and Records Administration. She also maintains CSTB's contact database; manages the content of the CSTB Web site; coordinates the layout and design

of *Update*, the CSTB newsletter; and designs book covers and promotional materials for outreach purposes. Prior to her move to Washington, D.C., Ms. Bishop worked for the City of Ithaca, New York, coordinating the Police Department's transition to an SQL-based time accrual and scheduling application. Her other work experience includes designing customized hospitality industry performance reports for a research firm, maintaining the police records database for the City of Ithaca, and freelance publication design. She is a visual artist working in oil and mixed media. She holds a B.F.A from Cornell University.