ELECTRONIC RECORDS ARCHIVES

**INTRODUCTION TO PRESERVATION AND ACCESS LEVELS
CONCEPTS
(PAL v1.0)
(WBS #1.1.15)**

for the

**NATIONAL ARCHIVES AND
RECORDS ADMINISTRATION**

**ELECTRONIC RECORDS ARCHIVES
PROGRAM MANAGEMENT OFFICE
(NARA ERA PMO)**

Final
December 5, 2003

FINAL

INTRODUCTION TO PRESERVATION AND ACCESS LEVELS CONCEPTS
Signature Page

Program Director,

I recommend approval of the Introduction to Preservation and Access Levels Concepts (PAL).

_____   _____

Dyung Le,
ERA Program

Approved,

_____   _____

Kenneth Thibodeau,
ERA Program Director

FINAL

Document Change Control Sheet

Document Title: Introduction to Preservation and Access Levels Concepts White Paper

Date	Filename/version #	Author	Revision Description
12/05/03	ERA.DC.PAL.1.0.DOC	Jon Valett/Ken Thibodeau	Final

FINAL
TABLE OF CONTENTS

List of Tables

FINAL

INTRODUCTION TO PRESERVATION AND ACCESS LEVELS CONCEPTS

1.0   Purpose

The purpose of the *Introduction to Preservation and Access Levels Concepts (PAL)* paper is to clarify what is involved in the preservation of electronic records as it relates to the Electronic Records Archive (ERA). The paper does not completely define this concept, nor does it attempt to fully articulate all its nuances. It provides background information to the *ERA Concept of Operations (ConOps)* and *ERA Requirements Document (RD)*.

1.1   Introduction

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

Given that both the purpose and the proof of preservation of electronic records is access, National Archives and Records Administration's (NARA's) goal is to preserve electronic records in a way which keeps all of their essential properties – as records – intact, when enabling users to access the records using state-of-the-art technology at any time in the future. This means being able to reproduce electronic records without dependence on specific hardware and software formats.

FINAL

1.2   ERA Program Overview

ERA will be a comprehensive, systematic, and dynamic means for preserving virtually any kind of electronic record, free from dependence on any specific hardware or software. The ERA, when operational, will make it easy for NARA customers to find records they want and easy for NARA to deliver those records in formats suited to customers' needs.

2.0   Document Assumptions

The following assumptions were made during the development of this document.

- The PAL's intended audience is the ERA development contractor and the ERA Program Management Office (PMO).
- The PAL is meant to provide guidance for the proposal phase of the ERA program. The concepts embodied in the PAL are expected to be further refined and explored during the Analysis and Design Phase of the ERA program.

3.0   Acronyms and Definitions

Table 3-1, Acronyms List, contains a list of acronyms used herein.

ACRONYM	DEFINITION
ConOps	Concept of Operations
COTS	Commercial Off the Shelf
ERA	Electronic Records Archives
GIS	Geographic Information System
ISO	International Organization Standardization
IT	Information Technology
NARA	National Archives and Records Administration
OAIS	Open Archival Information System
PAL	Introduction to Preservation and Access Levels Concepts
PAM	Persistent Archives Method
PMO	Program Management Office
POP	Persistent Object Preservation
RD	Requirements Document

Table 3-1: Acronyms List

4.0   Methods for Preservation

There are various strategies for digital preservation, including:

- Museum or Technology Preservation: maintaining the original hardware and software components in order to maintain the ability to reproduce the record

FINAL

- Format migration: converting all records from older to newer formats apace with the evolution of technology,
- Emulation: providing environments that recreate the environment in which the records were created, and
- Persistent Archives Method (PAM): keeping, and if necessary converting, records to a format that respects all its essential properties, but reduces or eliminates dependencies on specific hardware or software, enabling faithful reproduction over varieties and generations of technology.

The PAM technique (sometimes referred to as Persistent Object Preservation (POP)) has one key difference from the others: it focuses on preserving the specific properties of records, while the others focus on solving the problem of technology obsolescence.   However, the PAM technique is aligned with a major direction in the evolution of information technology: to enable business, government, education, research, and other activities conducted over the Internet, the information objects which are communicated in these activities must be basically independent of the technology infrastructure. Because of this mainstream direction, more and more products are appearing in the market that can be used to implement a PAM strategy, whereas the other techniques listed remain niche technologies. Thus, the PAM strategy appears the best option for satisfying NARA's mission requirements for preserving and providing access to electronic records.

5.0   Introduction to Persistent Archive Method (PAM)

PAM rests on two (2) basic assumptions: (1) the essential properties of a record can be specified, and (2) in many cases, a record may have several different, but equally valid materializations. The second assumption is implicit in many current applications. For example, many organizations create records using computer applications, such as word processing, but save these records printed on paper, and they may disseminate them electronically in HTML or '.pdf' formats. Similarly, a verbatim transcript of a hearing is commonly considered a reliable record of that hearing, even though it may have been recorded originally on an audio medium. Such practices presume that different materializations of the same record are equivalent for the purposes for which the records are created. PAM extends this presumption over time, but adds a requirement for the explicit specification of the essential properties of a record as a control which makes it possible to determine if a reproduction is authentic. Generically, NARA defines the essential properties of a record as including its content, structure, context, and presentation. Identifying essential properties does not necessarily entail that these properties cannot change, but that the variability of any property must be specified and not change over time. For example, one of the advantages of computer assisted design is that 3-dimensional digital designs can be rotated and viewed from different angles. The ability of a user to rotate the design defines a variability which must be preserved. If the design were saved in a format which only supported a static display, the result would be considered a version or extract of the original design, not an authentic copy.

PAM has the characteristics of being comprehensive (it should work for a wide variety of record types and digital formats), scalable to enormous quantities, infrastructure independent, and

FINAL

flexible. The PAM method is being developed within the context of an implementation of the International Organization Standardization (ISO) standard for Open Archival Information Systems (OAIS).

PAM is based on an object-oriented approach that is familiar to Information Technology (IT) professionals, but has only recently been applied to the work of archives and libraries. In the object-oriented view, all objects are characterized as having specific attributes and behaviors. Traditionally records have not been conceived as having behaviors, but in a digital environment, it is necessary to specify what operations or processes can be appropriately applied to any object, and what the results of such processing should be. For a textual record, for example, a required behavior is that its content must be presented in a page layout. Every record has a behavior of occupying a specific position in an original order imposed by the records creator, although in a digital context the order and position are most likely to be logical rather than physical. In the PAM approach, electronic documents, which have both content and form or structure, are represented as complex real-world objects having attributes and behaviors that must be maintained. More complex cases, such as in a decision support or Geographic Information System (GIS), the electronic records would be represented as complex objects that include descriptions of behaviors that explain how their subcomponents interact with other objects to form the records. For example, a cartographic record displayed in a GIS may include elements drawn from dozens of different layers in the system. The behaviors of such a record include all of the operations needed to retrieve the required data from the appropriate layers, defined in the logical data model, and cause them to be displayed correctly in a map or chart.

The context of a record is determined by its "provenance," an archival concept that can be defined loosely as the context in which a record was created, and the "original order," an archival concept designating the connections among records established by the records creator. An electronic record's provenance and original order are represented by defining groups of objects as logical collections. For example, in a contemporary office e-mail system, each message would correspond to an object with clearly-defined attributes (to, from, and date, among others), while the group of all messages from an organization could be represented as a collection, with the original order being the organization of messages in user accounts. In an object-oriented approach, a collection of objects can itself be represented as an object.

Representing and encapsulating complex records from contemporary information systems as objects and collections requires a flexible, open, standardized, self-describing and extensible modeling language that will preserve the content, context, and structure of records. Persistently preserving records would entail encoding of record attributes, behaviors, content, and structure into an encapsulated object. This object would then represent all information required to re-instantiate the record at a future time. A parallel approach would be implemented to re-instantiate aggregates of records, such as files and series. By expressing record properties in such formal models, PAM reduces hardware and software characteristics. Software mediators will enable future technologies to interpret the models and the metadata.

As more and more government transactions are conducted via e-Government, electronic records will increasingly be created in formats that either are, or can easily be converted to, persistent formats. In the near future, however, NARA will need to ingest and preserve large quantities of

FINAL

records in formats subject to rapid obsolescence. In order to preserve them, eventually NARA will need to transform records to persistent form using the open and standardized modeling/encoding language. ERA will need to implement tools that will enable NARA to transform records received in legacy formats into PAM formats. NARA will need different transformation tools specifically designed for different record and data types.

NARA will stipulate the attributes and behaviors of records that need to be preserved, and define related metadata. This policy will govern the transformation of records received in legacy formats and will also guide agencies in creating records in formats that conform to these specifications.

6.0   ERA Preservation Strategies

The breadth, diversity and complexity of the electronic records challenge require a comprehensive and coherent approach for preserving and providing sustained access to authentic electronic records. This overall approach specifies what will be done to preserve specific classes of records and types of aggregates of records, as well as types of digital data and files. The approach must include templates for modeling the properties of records and aggregates, along with methods for reproducing the records and arranging them properly relative to other records in the same aggregate. These methods will vary depending on the data types in which the records are encoded and the digital files in which they are stored. A preservation strategy may be defined as a specification of the methods which will be used to maintain, reproduce, and/or arrange a class of electronic records or record aggregates. Given the expectation of continuing change in information technology, preservation strategies will need to evolve over time. Preservation methods, then, include not only the methods for maintaining, reproducing and arranging records, but also methods for migrating or transforming the digital components of electronic records from one format to another.   Given the magnitude of the challenge NARA faces, there is a premium on adopting preservation strategies which are consistent among themselves, minimize the variety of methods NARA needs to employ at any given time, and reduce the need for changing methods over time. The hierarchical approach proposed for templates in the paper, "*Introduction to Policies, Templates, and Requirements Concepts*," outlines a framework for developing a coherent and manageable set of preservation strategies.

6.1   Preservation Strategies for Aggregates

An archival aggregate of records is an ordered set. Aggregates may have hierarchical levels. ERA will need to have a way to rebuild the structure of any aggregate of electronic records, and to place the records in the arrangement defined for that aggregate. The template for an aggregate must specify its structure, enable members of the aggregate to be identified, and determine how members are related to one another within the arrangement defined for the aggregate. This template would map to methods ERA would apply to identify an aggregate of records, build the structure of the aggregate, identify the members of the aggregate, and retrieve and place each member properly in the structure. For example, for records organized according to a traditional file classification plan, ERA should be able to establish a hierarchical structure corresponding to

♦ **National Archives and Records Administration** ♦

FINAL

the file plan, place any retrieved file in its proper place in that structure, and place all the records that belong in that file in their proper order.

In addition to the aggregates established by the records creator, NARA may define sets of records to facilitate execution of its lifecycle management functions. Such 'Lifecycle Management Sets' are established by management decisions on how to meet generic, and usually ongoing, requirements. They are usually independent of provenance. For example, a policy to retain that all permanent electronic records in the formats in which they were transferred to ERA, as well as any migrated or persistent format, defines the Lifecycle Management Set of all files in native formats. Whether a preservation strategy is needed for a given Lifecycle Management Set depends on the nature of the set. The set of all files in native formats is risk mitigation measure to enable alternative migrations or transformations from the original received files. To be effective it would require a preservation strategy which specified a method for reading each of the formats included in the set.

6.2   Record Level Preservation Strategy

An electronic record is encoded in at least one digital object. For example, a textual document may be recorded in a word processing file, and a transaction, such as a retail sale, might be recorded in a row or tuple in a data base. But the word processing file might contain a digital photograph. The photograph could not be displayed using only word processing software. Similarly, viewing the sales transaction might entail putting the data into an online form. The form might be in HTML. In each of these cases, the electronic record would comprise two (2) digital objects. A digital object which is needed to reproduce an electronic record is called a digital component of that record. The preservation strategy for all electronic records requires that all the digital components of each record be identified and the format of each component be specified. The strategy would stipulate the methods to be used to assemble, organize and present the components in order to reproduce the record.

7.0   Preservation and Access Levels (PAL)

Preserving electronic records in a way which both satisfies records management requirements and makes them relatively immune to changes in information technology infrastructure is an ideal. Achieving this ideal will be complex and costly. Therefore, NARA is developing a stratified approach which provides continuing access to the records while providing flexibility in managing preservation over time. This approach predicates that there will be a minimal level of preservation and access services provided for all permanent electronic records brought into the system, and that this level may be increased over time, moving towards the goal of persistent preservation described above. This section presents examples of potential levels of preservation and access within ERA. The preservation and access level for any record or set of records will be specified in a Preservation and Access Plan developed at the time the records are scheduled for transfer to ERA. Three (3) preservation and access levels have been proposed. Alternative preservation and access levels may be developed to accommodate changing technology or improved preservation techniques.

FINAL

7.1   Basic Preservation and Access

At a basic preservation and access level, ERA will preserve all records in their original format, encapsulating them with basic archival information describing the record, including such items as date received, agency, and series, as well as, the hardware and software characteristics that will need to be known for later transformation. The only guaranteed access service for records at this level will be the delivery of the files in their original format. For some subset of these records, such as for data types that only require software viewers, NARA may also have tools that enable retrieval of the record in its native format. This level will also include the ability to identify and retrieve all the records in any aggregate of records. For some subset, such as records arranged only according to a DOS directory path, it may also be possible to instantiate the arrangement. Electronic records received in legacy formats may not be converted to an enhanced or PAM format for some time because of technical difficulties or costs of such conversion. The basic services will be the only ones available for such records.

Temporary and non-accessioned records that are only kept within ERA for a few years may never move from this minimal level of preservation.

If the stored format becomes obsolete and there is (expected) demand for access to the records, it may be necessary to migrate the records to a current format, or convert them to a persistent format, to satisfy the demand. The option between a current format and a persistent format should be decided on the basis of relative costs and difficulty of the transformation, and whether archival requirements for a persistent format have been adequately defined. These options define two (2) other Preservation and Access Levels.

7.2   Enhanced Preservation and Access

There will undoubtedly be many cases where the file copy or software viewers available at the basic level are not adequate, but where there are also barriers preventing conversion to a persistent format.   In such cases, NARA would provide an enhancement over the basic preservation and access service in order to satisfy requirements for continuing access.

The enhanced preservation and access level might include additional external metadata indicating attributes defined in NARA standard template for a given record type or aggregate (e.g., title, date, issuing office or official for formal reports).

Enhanced service might be achieved by migrating records from obsolete formats to newer ones for which better access software is available. For example, a record might be converted from a proprietary format, which required specific hardware or software, to a standard format which is both defined in an open standard and has sufficient market penetration and support to justify an assumption that it will remain possible to access the record using Commercial Off the Shelf (COTS) products for many years.

♦ National Archives and Records Administration ♦

FINAL

7.3   Optimal Preservation and Access (PAM)

The optimal preservation and access level for archival records would include retaining the records in the formats in which they were received, and in a persistent version which satisfies requirements for the preservation of authentic records and facilitates access using advanced technologies. The persistent version is expected to be an implementation of the Persistent Archives Method. The persistent version would explicitly indicate the elements of content and structure, presentation and behavior which have to be preserved. That is, it would apply the NARA standard template for the record or aggregate type to the stored digital components. The encapsulation or markup of the records would conform to open, non-proprietary standards to achieve maximum feasible independence from specific information technology infrastructure.

Note that persistent versions could take different formats. For example, for textual records received in a character-based format, where there was a variety of record types, the content covered a variety of topics, and NARA anticipated requirements for rich content searching, the appropriate persistent format would involve semantic and syntactic markup. But if textual documents were received in digital image format and the anticipated demand was for copies of specific documents presented with their original 'look and feel,' the appropriate persistent version might be an image format encapsulated in metadata which would support expected search strategies.

Conversion to persistent archival format will be more expensive than simple encapsulation or migration to enhanced formats; however, neither of the other approaches provides any assurance of permanent preservation or accessibility. PAM will ensure that the record will be accessible for the longest period of time. A PAM strategy, as described above, will also reduce the technical variety that NARA must support over the long term. Given the expectation of continuing change in IT, it is likely that at some future time, NARA may decide to transform preserved records to newer persistent formats, but such conversions are likely to be much less frequent than format migrations. Moreover, NARA will probably have greater flexibility in deciding when to execute the transformations. While format migrations will be driven by the obsolescence of current formats, transformations between persistent formats will likely be done to take advantage of improved technologies. NARA would have the option of performing such transformations opportunistically, such as when records are retrieved in response to reference requests or when storage subsystems are replaced as part of routine technology refreshment. Such considerations should make the long-term costs of the PAM strategy less expensive than alternatives.

8.0   Relating Preservation Strategies to Disposition Agreements

A disposition agreement which provides for transfer of electronic records to NARA entails requirements for preserving and providing access to these records after transfer. How NARA will meet these requirements should be set out in a Preservation and Access Plan for the records covered by the agreement. The Preservation and Access Plan should identify the types of records, record aggregates, and the data types of their digital components. It should specify the application of the current NARA preservation strategies for these types of objects. However,

♦ **National Archives and Records Administration** ♦

FINAL

given the great variation in the way electronic records are created and retained, there may be no established preservation strategies which are adequate or appropriate for the records covered by an agreement. In such cases, the representative of the records creator, the NARA appraiser and a NARA preservation specialist should work together to develop a suitable Preservation and Access Plan.

9.0   Conclusion

This paper has presented an introduction to the preservation of electronic records in the context of ERA. It states the purpose of preserving records, rather than digital artifacts, and describes how the preservation process for electronic records extends through the production of copies of those records. It sketches an idealized vision of a Persistent Archives Method, but presents a multifaceted and flexible approach to realizing that vision. It shows how that approach can be integrated with the lifecycle management of records.