SLAC- PUB- 7811
January 2000

WRIT IN WATER?
An exploration of the gap between Archival construct and practice in the machine-readable environment

Jean Marie Deken

Stanford Linear Accelerator Center
Stanford University
Stanford, California 94309 USA

Abstract

The late twentieth century explosion of electronic record-keeping media poses new and compelling challenges. These new challenges arise from the increasing complexity of the technology of documentation, and from the increasing fragility of that technology over time. Challenges posed by changes in the collection, dissemination and storage of information and records are covered, and an overview of currently proposed approaches to appraising, storing and preserving the new media is provided.

Invited talk presented at the
WORKING WITH KNOWLEDGE CONFERENCE
5/6/98 --5/7/98, Canberra, Australia

Work supported by Department of Energy contract DE-AC03-76SF00515

Introduction

So sure was he that nothing would remain of his life's efforts, the English poet John Keats (1795-1821), in a paroxysm of romantic despair, chose the epitaph: "Here lies one whose name was writ in water."[1] And yet, more than 170 years after his death, his works are still with us, still read, studied, admired and imitated. Luckily for Keats, his life's work was recorded on paper, the dominant record-keeping medium of the day, and a fairly stable long-term storage instrument.

In the late twentieth century, we live on a planet where there has been an explosion of electronic record-keeping media, and, as a result, many intellectual achievements of our age may prove to truly be "writ in water," disappearing without a trace soon after they are created. The records of late twentieth-century science, in particular, pose new and compelling challenges to the scientists, archivists and historians responsible for preserving and interpreting them. These new challenges arise from the increasing complexity of the technology of documentation, and from the increasing fragility of that technology over time. Moreover, the challenges posed by changes in the collection, dissemination and storage of information are so profound and far-reaching that the archival profession itself seems poised on the edge of radical change and redefinition as a result.

Overview and Definitions

To adequately assess how much things have changed, it is useful to first establish how things "used to be." For the past 56 years, archival administration and records management in the United States and abroad have been based, in large part, upon a fundamental construct called life-cycle management. This construct, formulated in 1940 by appraisal archivists with the United States National Archives (Brooks, p. 5) holds that all records pass through the same life cycle of: creation, use, storage and disposition. (Figure 1)

Disposition in the life-cycle of records is the point at which a record is either officially declared permanent, because of its historical value, or is officially disposed because it has ceased to have any historical or informational significance. The process of determining the appropriate disposition of a record as either permanent or temporary is called appraisal and scheduling**.**

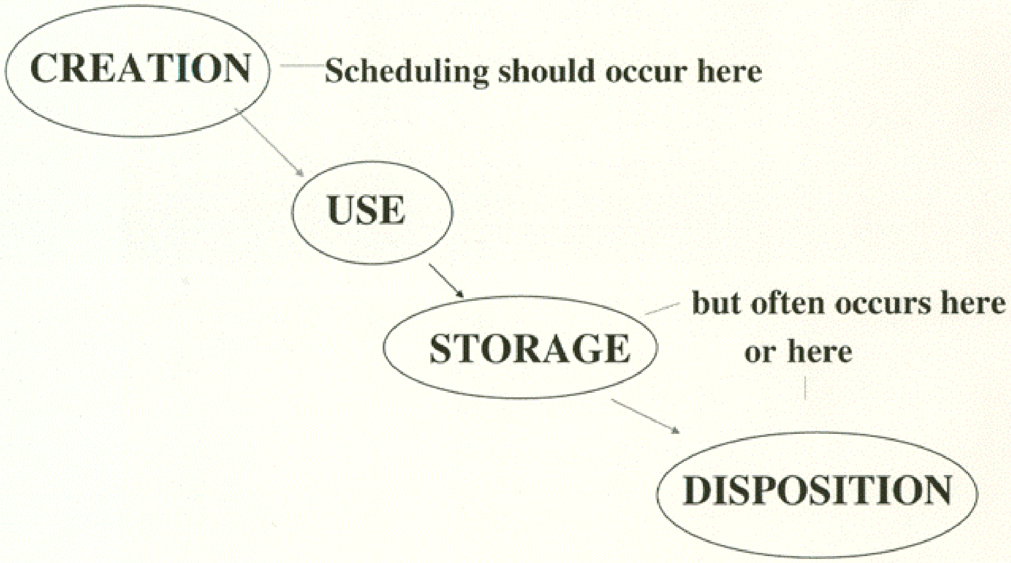

Figure 1: Scheduling Of Paper-Type, Human-Eye-Readable Records And Information

The Recent Past

When this life cycle of records construct was first articulated in the US, most of the data and information which passed through the life-cycle steps were paper-type[2] and most were also human-eye readable.[3] In addition, most information passed consecutively through the life-cycle, that is, creation was followed by a period of active use, which was then followed by a period of storage, which was then followed by disposition.

Accepted archival theory held that the scheduling of records, that is, the determination of ultimate disposition, should occur as early in the life-cycle of the records as possible, preferably at the time of their creation (Brooks, p.1-2). Common archival practice, however -- because of a variety of factors -- has often resulted in the scheduling of the records not at the point of creation, not at the point of active use, but at the time they are retired to storage, IF the storage location is controlled by an archivist or records manager. If not, scheduling typically occurs at the time of disposition.

A second fundamental construct of archival management, also developed in the environment of paper-type, human-eye-readable, records, is the principle of physical transfer**.** This principle holds that records that are scheduled for permanent retention must be transferred to the physical custody of an archival repository, where they will be arranged and described to facilitate their continued use, and properly stored to facilitate their continued survival. In actual practice, however, the principle of physical transfer -- much like the principle of scheduling -- has often been abridged or ignored by records creators and users. Massive quantities of ultimately disposable paper-type records with admittedly short-term value have routinely been transferred to archival repositories, while creating agents and agencies have been notoriously reluctant to part with physical custody of even small volumes of paper-type records which they consider to be historically significant..

Such has been the status of both archival practice and archival holdings in the United States from roughly the end of World War II up to the 1965 publication of T. R. Schellenberg's landmark of archival literature, the monograph The Management of Archives (Schellenberg). Since 1965, however, practically every element in the archival construct of life-cycle management has undergone significant change.

Increasing Complexity of Documentation Technology

Today, in May 1998, a staggering amount of the documentation, information, and data created in the course of human endeavor is not paper-type, and is not human-eye readable. This shift has occurred so rapidly and across so many diverse disciplines that it has even destabilized the basic vocabulary which had been used heretofore to describe and discuss the creation, use, transmission and storage of information. The words record, documentation, and archive, for example, mean very different things to archivists than they do to Information Resource Management (IRM) professionals. Archives, the noun, became archive, a verb, upon its relatively-recent migration from the vocabulary of archivists into that of computer programmers and operators, and it means significantly different, although somewhat similar things, to both professions..

Late-twentieth-century records which are not paper-type include those which store information on photographic film, magnetic media, and optical media. Present-day records that are not human-eye-readable must be mediated or translated for the human eye (and ear) by some type of computer software and/or hardware. Two terms that have been used by archivists to describe these records are electronic records and machine-readable records. I prefer the latter term as both more inclusive and more accurately descriptive.

Machine-readable records are archivally quite different from human-eye-readable records in a number of significant ways. They are much more compact, they are easier to create, alter, and transmit; and they are, in some ways, easier to store. Some types of machine-readable records have the same life-cycle as that of paper-type records, but, unlike paper-type records, they tend to pass through some stages of their life cycle simultaneously, rather than consecutively. (Figure 2)

For most machine-readable records, the middle two phases of the life-cycle tend to be conflated: since creation, revision and transmission are so flexible and inexpensive in the machine environment, machine-readable information tends to have a longer "use" phase, and, since storage of machine-readable information is so compact and so easy to accomplish, "storage" tends to occur at the same time and in the same location as "use."

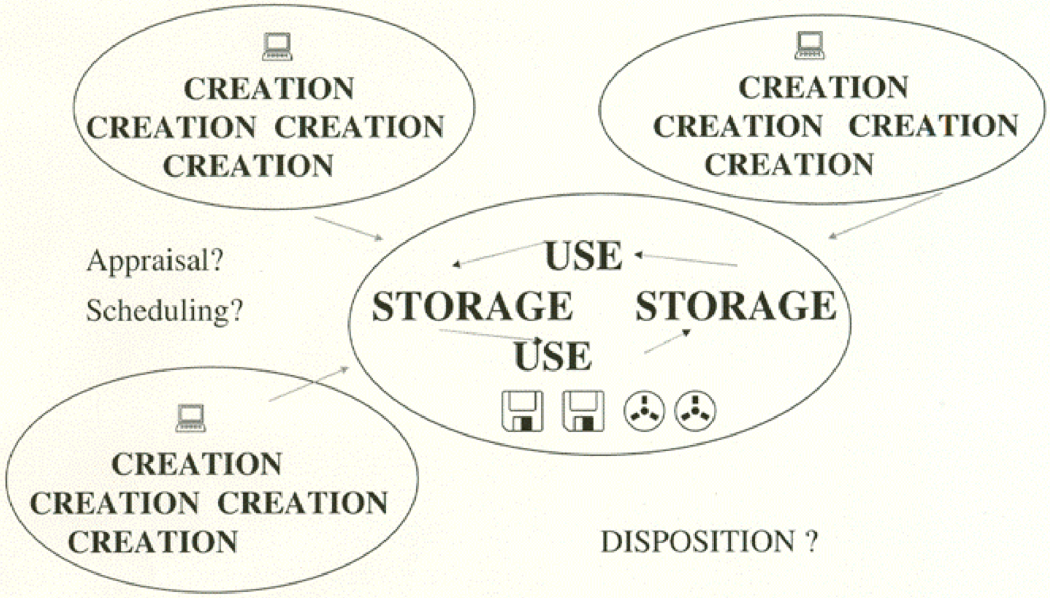

Figure 2: Life - Cycle Construct Applied To Machine - Readable Records

This latter fact, the tendency of the use and storage phases of machine-readable data to occur simultaneously and in the same location, has had significant negative impact on the archival principle of physical transfer. And, since the scheduling of records to determine their historical significance and ultimate disposition has typically taken place at the time of physical transfer into storage (even though it was supposed to occur much earlier in the life-cycle), in this brave new world of machine-readable records, where there is no physical transfer of records, there tends to be no archival appraisal and, consequently, no scheduling.

The absence of archival appraisal and scheduling in the machine-readable records environment in and of itself could be a relatively easily-correctable problem, except that it is compounded by another important attribute of the machine-readable record environment: inherent instability.

Stability of Old Documentation Technology

The best way to demonstrate the instability of the machine-readable environment is by comparison to the relative stability of the prior, human-eye-readable record environment. A high-profile example of human-eye-readable documentation is the 220-year old U. S. *Declaration of Independence*.[4] The life-cycle of this document began with the collaborative creation by the Continental Congress, occurring in the city of Philadelphia, Pennsylvania, over the summer months of the year 1776. Creation of the *Declaration* was then followed by its official adoption by the Continental Congress on July 4, 1776; its approval by all of the thirteen colonies by July 9, 1776; and the signature of an official copy "engrossed on parchment" by the delegates to the Continental Congress on August 2, 1776.[5] The official, signed, parchment *Declaration* was in use by the Continental and Confederation Congresses through 1789, at which time it began the storage phase of its life-cycle. (Figure 3)

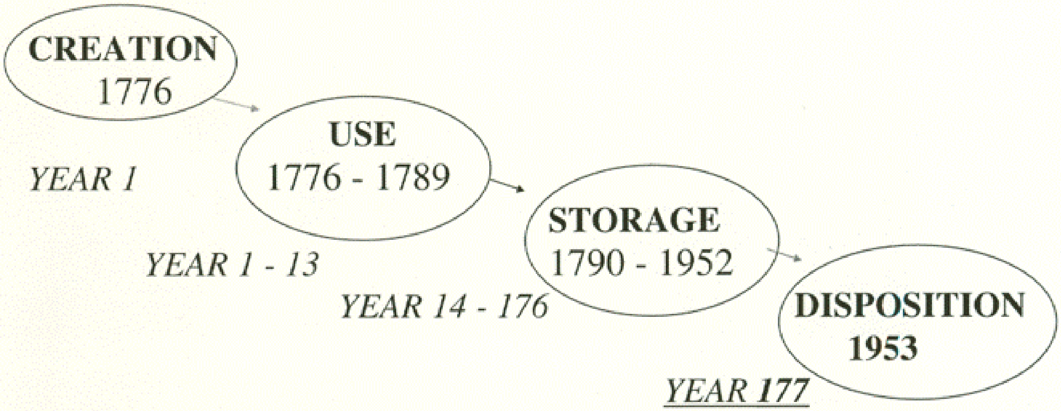

Figure 3: Scheduling and Disposition of US Declaration of Independence

While there was never any question in the mind of any official of the newly-minted republic of the United States of America that the disposition of the official copy of the *Declaration* was to be permanent, the storage location of the official document has shifted a total of 28 times so far in its 220 years of existence. Furthermore, although contemporary commentary indicated that there was noticeable deterioration of the condition and readability of the *Declaration* as early as 40 years into its existence, it was only upon the 100th anniversary of its approval, in 1876, that serious, "official" attention began to be paid to arresting and possibly reversing its deteriorating physical condition. The United States' National Academy of Sciences studied the *Declaration in* from 1880-81, and again in 1903. These studies led to its temporary removal from public display in 1904 in order to limit its exposure to the deteriorating effects of excess light and humidity until appropriate protections could be devised. Further and increasingly more elaborate steps have been taken to adequately store and preserve the *Declaration* since 1904.

In 1953, the *Declaration* was transferred to the US National Archives, beginning the "disposition" phase of its long existence. It is now enshrined in a specially-designed and monitored exhibition case, within a specially designed exhibition hall in the National Archives building in downtown Washington DC.

Instability of New Documentation Technology

A comparison of the life-cycle of this well-known paper-type, human-eye readable document with what is known about the life-cycle of present-day machine-readable records will amply demonstrate the sweeping changes, which have occurred, and are continuing to occur, in the archival landscape.

I will start by making some generalizations about my understanding of the status of machine-readable records. Machine-readable records are being created by more and more people than ever before; they are being used by more people, and their use is being shared simultaneously by people in separate locations. Machine-readable records are also being stored in what is sometimes called a distributed environment, which means not necessarily all together, and also not necessarily, nor usually, at the same geographic location as the users. In effect, almost everything about the life-cycle of machine-readable records is distributed: distributed creation; distributed use; and distributed storage. However, disposition in a distributed environment appears to often be premature and uncontrolled because it has not yet found a natural, logical and controlled place in the flow of information.

In a machine-readable records environment, more than ever before, appraisal and scheduling must occur in as close a proximity to creation as possible. The life-cycle of the machine-readable record can be painfully short because of the short life-expectancy of its two component parts: the machine AND the "reading" program that translates what is stored into something comprehensible to the human eye. Exactly how short is "painfully short?" The growing consensus seems to be two to five years, or, put another way: "Digital information lasts forever, or five years, whichever comes first."[6]

Backward compatibility of software beyond the latest one or two versions has not yet been a priority of the off-the-shelf software industry; and the track record of in-house programmers in this area tends to be just as bad. As a US National Research Council study states: "the greatest barrier to contemporary and future use of scientific data by other researchers, policy-makers, educators and the general public is lack of adequate documentation," and "a general problem prevalent among all scientific disciplines is the low priority attached to data management and preservation by most agencies. Experience indicates that new research projects tend to get much more attention than the handling of data from old ones, even though the payoff from optimal utilization of existing data may be greater. (NCR p. 2)

Recent reports by both the Research Libraries Group and the National Research Council cite several instances of important US documentation that has been threatened with premature extinction due to its machine-readable format. The RLG report details, for example, the measures taken to preserve the 1960 U. S. Census from loss:

> As it compiled the decennial census in the early sixties, the Census Bureau retained records for its own use in what it regarded as "permanent" storage. In 1976, the National Archives identified seven series of aggregated data

> from the 1960 Census files as having long-term historical value. A large portion of the selected records, however, resided on tapes that the Bureau could read only with a UNIVAC type II-A tape drive. By the mid-seventies, that particular tape drive was long obsolete, and the Census Bureau faced a significant engineering challenge in preserving the data from the UNIVAC type II-A tapes. By 1979, the Bureau had successfully copied onto industry-standard tapes nearly all the data judged then to have long-term value. (RLG p. 2)

Here is a modern, machine-readable set of documents, or more correctly, group of datasets, that the creating agency appraised as permanent at the time of creation -- note that the appraisal took place at the correct point for appraisal under the original archival construct. A short 16 years later, however, when a stake-holder inquired about the "permanent" records, they were discovered to be in need of rescue. The rescue effort took three years, and not all of the data could be saved. (Figure 4)

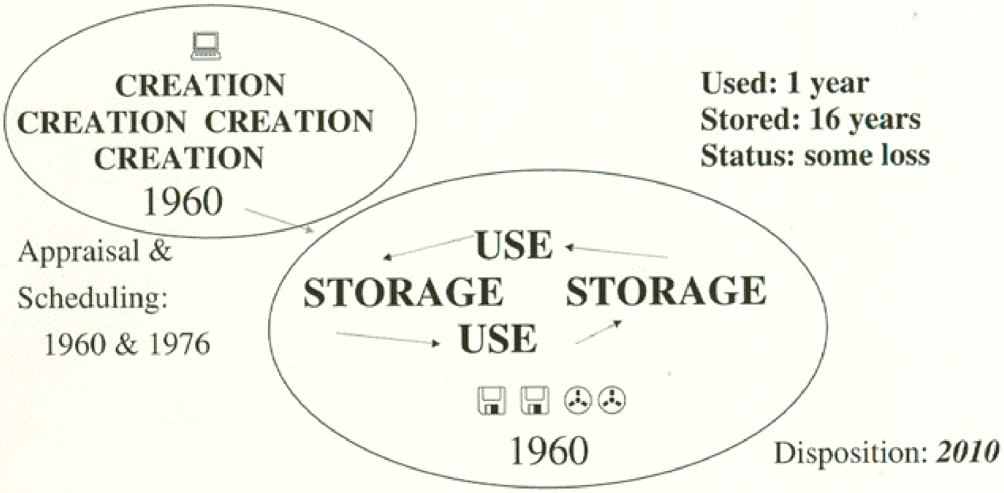

Figure 4: Scheduling and Disposition of US 1960 Census

Other historically significant records have disappeared completely, including the first e-mail message (the sender of which can not be determined because the 1964 message is gone), and U. S. satellite observational data of the Amazon basin in Brazil in the 1970's (RLG p. 3, NRC p. 31), and many of the original pages of the World Wide Web. Whereas the "window of opportunity" to appraise and preserve a certain eighteenth-century parchment human-eye-readable document appears to have been equal to or greater than 150 years, the "window of opportunity" for machine-readable information, some of which is just as important, far-reaching, and life-enhancing as that treasured but admittedly ill-kept document, is substantially less than a decade, and may even be shrinking.

Viability of the "Life-Cycle" Construct for Machine-Readable Records

It may no longer be helpful to view the functions of creation, use, storage and disposition as a cycle, because that construct is based on the assumption that the elements of the cycle are sequential, rather than concurrent. The European Communities DLM-Forum on Electronic Records has proposed a revised life-cycle for digital information, positing that it travels through three phases: design, creation and maintenance (EC 3.1). While this newer construct may, in fact, be an improvement (so long as archival appraisal is understood to be an integral part of the "design phase") it shares a serious limitation with the life-cycle construct in that both support the illusion that there exists a span of time after the generation of information during which deliberations can be made about whether or not the information should be permanently retained. In the machine-readable environment, this illusion leads to slow action or inaction, both of which lead, inevitably, to loss.

An additional problem that faces archivists operating in the machine-readable environment is the problem of the "record-ness" of machine readable information, which is a problem involving two aspects of content, integrity and "fixity" (RLG 14). Establishing the "record-ness" of human-eye-readable entities or documents is routine in the twentieth century, because such documents are relatively stable and because conventions of integrity and fixity have been long established for them. We tend to think of these conventions as integral to human-eye-readable information, but they have, in fact, been artificially constructed over centuries of negotiation, practice, and habit. Establishing similar integrity and fixity for machine-readable records may mean adopting the practice of publishing a version of the digital information in an alternate medium, or of capturing a digital snapshot of data at a set point in time. It will likely also involve establishing conventions for the consistent and persistent naming of digital objects over time (Payette). The demands of integrity and fixity for records may also require us to recognize that the machine-readable environment does not naturally or automatically create records -- that it is a "meta-medium," a set of layered services built from flexible elements (Agre) -- and that the "record-ness" of digital objects is something that must be both consciously generated and constructively protected.

Automating appraisal

In an era when so much information is "born digital," perhaps it is time to introduce a new archival construct: that of the record that is "born appraised," or "born scheduled." Several studies now underway are exploring the notion of automating the appraisal and transfer of machine-readable information by building these functions into application software (DoD, AHDS 1.4).

Even when provisions are made to allow appraisal and scheduling of machine-readable records to occur when they are supposed to occur -- at the time of creation of the record--

the life expectancy of information in the new machine-readable environment is too short to allow archivists to continue to depend upon the relatively passive transfer and disposition methods currently in use, i.e.. the archives waits for the creators and/or users to initiate transfer of important information to the archives.

Transforming transfer

One new approach being tried is the acceptance and institutionalization of systems of distributed archives. (Figure 5) Both the NRC and the RLG reports recommend and even encourage adoption of a system of distributed archives as the only economically feasible and scientifically sound approach to controlling and preserving the massive machine-readable scientific data systems now being created.

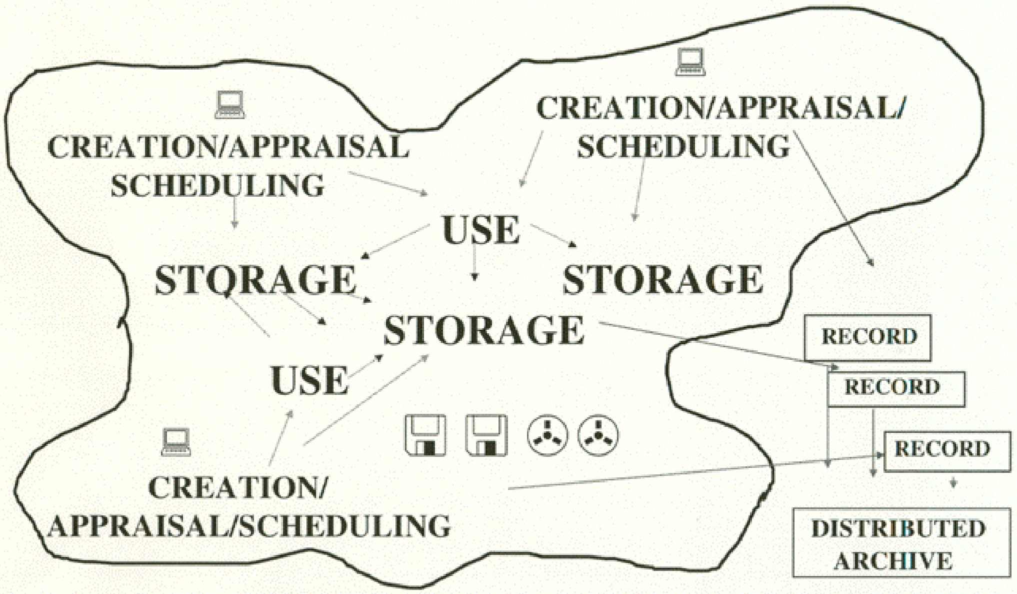

Figure 5: Automating Appraisal & Transforming Transfer

While the distributed archives approach effectively deals with the transfer and storage phases of the life-cycle of machine readable records, it does not adequately address the disposition phase: distributing the archives by allowing them to remain in the hands of the creating persons or agencies does nothing to address issues surrounding the life-expectancy of the hardware, software and storage media.

In the machine-readable information environment, effective preservation plans must include realistic assessments of costs. Once historically-significant information has been identified, and its “record-ness established, the best permanent-storage medium for the records must also be identified. In some instances, this may be the original medium, but when the original medium is machine-readable and, therefore, technology-dependent, careful consideration must be given to the continued costs of repeatedly migrating the information to new technology at regular 5-year intervals in perpetuity. Data migration will always be expensive, even as information storage capacities rise and information storage costs decline, because migration involves personnel costs and opportunity costs for organizations that have not yet been sufficiently explored.[7]

However, the RLG Task Force on Archiving Digital Information, and other professional work groups are actively examining the fiscal and staffing implications of the brief life-expectancy of electronic hardware and software. An approach to preserving machine-readable data that was adopted by archivists quite early in this effort is a method called “technology refreshing,” which is simply periodically copying existing machine-readable information onto new media. The RLG Task Force has closely examined “technology refreshing” and has found it to be inadequate. RLG recommends a more sophisticated approach called “data migration.”

Data migration, “is a set of organized tasks designed to achieve the periodic transfer of digital materials from one hardware/software configuration to another, or from one generation of computer technology to a subsequent generation. The purpose of migration is to retain the ability to display, retrieve, manipulate and use digital information in the face of constantly changing technology (RLG p. 6). This migration process has been undertaken routinely in data processing departments of large organizations, for the migration of current data. The innovation of the RLG proposal is that it applies data migration strategies and procedures to information that is to be kept forever.

Both the RLG Task Force and the NRC Report recommend that a network of “specialized data centers” (NRC p. 48) or “certified digital archives” (RLG p. 38) be established in the United States. The RLG notes that while

> the first line of defense against loss of valuable digital information rests with the creators, providers and owners of digital information...Long-term preservation of digital information on a scale adequate for the demands of future research and scholarship will require a deep infrastructure capable of supporting a distributed system of digital archives.

The RLG Task Force further recommends an admittedly radical approach to machine-readable records and the problem of data-migration by proposing that

> digital archives may invoke a fail-safe mechanism to protect culturally valuable information.... Such a mechanism, supported by organizational will, economic means, and legal right, would enable a certified archival repository to exercise an aggressive rescue function... toward digital information that is in jeopardy of destruction, neglect or abandonment by its current custodian. (RLG p.8, 22, 40)

Conclusion

Whether we do away entirely with the life-cycle construct, or re-invent it to include "digital object identifiers," "distributed archives," and "data migration," the archival profession needs to move quickly and decisively to develop and implement an entirely new or radically revised construct, precisely because machine-readable information of historic significance is both precious and highly perishable, and it is perishing right now. Whether exercising an "aggressive rescue function" for machine-readable records, for example, is a realistic and achievable role for archives remains to be determined, but the idea is an excellent example of the bold new thinking and new constructs that I am convinced are now required of archivists if we are to preserve and curate a history more lasting than one that is "writ on water."

Acknowledgements

I wish to thank Gavan McCarthy, Director, Australian Science Archives Project, for sponsoring the *Working With Knowledge Conference*, and for inviting me to present this paper there. I am likewise grateful to Pat Kreitz, and David Leith, of SLAC, for supporting and encouraging my participation in the Conference; and to Barbara Cytowicz, ASAP Project Archivist and Conference Organizer, for an providing an outstanding conference experience to all, presenters and attendees alike.

REFERENCES:

Agre, Phil. *The Internet and Public Discourse.* First Monday 3 (3) March 2, 1998 (http://firstmonday.dk)

Arts and Humanities Data Service. Digital Collections: A strategic policy framework for creating and preserving digital resources. Version 3.1, April 24, 1998. First Public Consultation and Review Draft (http://ahds.ac.uk/manage/framework.htm#sec0)

Brooks, Philip C., Jr. *The Life Cycle Concept and The Development of Federal Records Centers*, NARA 95-212, The Office of Federal Records Centers, Directions for the Future (Report) Appendix 1, March 23, 1995, Washington DC

Commission on Preservation and Access and The Research Libraries Group. Preserving Digital Information: Report of the Task Force on Archiving of Digital Information, May 1, 1996.

Department of Defense Design Criteria Standard for Electronic Records Management Software Applications, DoD 5015.2-STD

European Communities DLM-Forum Electronic Records. Guidelines on best practices for using electronic information: How to deal with machine-readable data and electronic documents. Luxembourg: Office for Official Publications of the European Communities, 1997. (http://www2.echo.lu/dlm/en/gdlines.html)

National Research Council (1995) Preserving Scientific Data on Our Physical Universe: A New Strategy for Archiving the Nations' Scientific Information Resources. Washington, DC: National Academy Press.

This is a U. S. National Research Council (NRC) report on the proceedings and conclusions of the NRC Commission on Physical Sciences, Mathematics, and Applications Steering Committee for the Study on the Long-term Retention of Selected Scientific and Technical Records of the Federal Government. National Academy Press, Washington DC 1995. The study was "performed at the request of the National Archives and Records Administration (NARA), and partially supported by the National Oceanic and Atmospheric Administration (NOAA) and the National Aeronautics and Space Administration (NASA)."

Payette, Sandra. *Persistent Identifiers on the Digital Terrain.* RLG DigiNews: Volume 2, Number 2

Schellenberg, Theodore R., The Management of Archives. New York, Columbia University Press, 1965.

Society of American Archivists. *The American Archivist, Special Issue: 2020 Vision.* Volume 57, Number 1, Winter 1994.

Issue guest edited by Margaret Hedstrom, with contributions from David Bearman, Luciana Duranti, Joan Warnow-Blewett, and many other leading-edge archivists. (Note: Joan Warnow-Blewett, of the American Institute of Physics, was also a member of the NRC study team.)

NOTES:

[1] Rome, Italy, Protestant Cemetery. Entire epitaph is: "This grave contains all that was mortal of a young English poet who on his death bed in the bitterness of his heart at the malicious power of his enemies desired these words engraved on his tombstone "here lies one whose name was writ in water" John Keats. (http://www/members.aol.com/WordPlays/graves.html)

[2] I am using the term "paper-type" to denote fiber (paper, papyrus, etc.) as well as skin (parchment, vellum) writing and printing media, because my emphasis is on the physical characteristics of the media as they relate to storing and accessing the information, rather than the organic or chemical attributes of the media.

[3] There have been some paper-type records which were NOT human-eye-readable, for example, paper data punch cards and paper punch tapes. Although interesting for several reasons, these media typically have served as processing instruments rather than records, and hence, are not addressed by this paper.

[4] Information about the history of the US Declaration of Independence from the National Archives and Records Administration's "On-line Exhibit Hall" exhibition "The Declaration of Independence" (http://www.nara.gov/exhall/charters/declaration/decmain.html) and "The Declaration of Independence: A History (.../declaration/dechist.html)

[5] ibid. Not all delegates who signed the Declaration were present on August 2, 1776.

[6] RLG Report, page 5. Quote is attributed to Jeff Rothenberg, in Stepanek, Marcia "From Digits to Dust" *Business Week* April 20, 1998 p. 129. In a previous article (Rothenberg, Jeff. "Ensuring the Longevity of Digital Documents" *Scientific American*, January, 1995. p. 44) Rothenberg estimates the longest-lived of the machine-readable media, optical disk, at 10 years, based on media lifetime alone (not mediating software or hardware).

[7] "Data creators who attach little or no value to the long-term preservation of the data resources they create are unlikely to adopt standards and practices, which will facilitate their preservation. This is particularly true where those standards and practices are different from or more costly to implement than those which promise the cost-effective development of a data resource capable of fulfilling its intended use. Accordingly,... awareness-raising...needs to be addressed toward data creators in a manner which appeals to their interests." AHDS 1.2