

**Carrying Authentic, Understandable and Usable Digital Records
Through Time**

**Report
To the Dutch National Archives
And Ministry of the Interior**

by

**Jeff Rothenberg
Tora Bikson**

RAND-Europe

August 6, 1999

ACKNOWLEDGEMENT

Many of the archival and recordkeeping insights in this report are based on conversations with Hans Hofman, of the Dutch National Archives. Since some of the ideas expressed in these interactions may not yet have been published elsewhere, specific citations are not given, but these ideas represent a primary source of intellectual input to this report.

ACKNOWLEDGEMENT	i
1. Overview	1
2. Motivation	1
3. Scope of the study	3
4. Summary of results and recommendations	5
4.1 Products of the study	5
4.2 Observations and recommendations.	6
4.3 Proposed next steps	8
Annex A: A strategy for preserving digital records	10
A.1 The top-down flow of the strategy	11
A.1.1 Analyze the functions that records must support	11
A.1.2 Derive authenticity criteria	12
A.1.3 Decide which attributes of records must be preserved	13
A.2 The bottom-up flow of the strategy.	14
A.2.1 Analyze technological alternatives for preserving digital records	14
A.2.2 Choose a technological preservation approach.	15
A.3 Integration of digital and non-digital archives	16
Annex B: A framework for preserving digital records	17
B.1 A generic, parameterized digital record preservation process.	18
B.1.1 The creation of a digital record	19
B.1.2 Entering the digital record preservation process.	21
B.1.3 Extracting a record	22
B.1.4 Ingesting a record	23
B.1.5 Preparing a record for preservation.	24
B.1.6 Acceptance testing of records	25
B.1.7 Metadata	26
B.1.8 Providing access to preserved digital records	28
B.1.9 Permanence of the preservation process itself	29
B.2 Supporting infrastructure and repository.	30
B.2.1 Processing	31
B.2.2 Storage (repository).	32

B.2.3 Communications	33
B.2.4 Software	33
B.2.5 Personnel	34
B.2.6 Administration	34
B.3 Experimentation and prototyping process	35
Annex C: Testbed	37
C.1 Rationale	37
C.2 Research Questions	38
C.3 Scope	40
C.4 Tasks	41
Task 1	42
Task 2	44
Task 3	46
Task 4	48
Task 5	49
Task 6	54
Task 7	54
C.5 Results	55
C.6 Testbed infrastructure requirements	55
C.6.1 Testbed processing requirements	56
C.6.2 Testbed storage requirements	56
C.6.3 Testbed communications requirements	57
C.6.4 Testbed software requirements	57
C.6.5 Testbed personnel requirements	57
C.6.6 Testbed administration requirements	58
C.6.7 Testbed start-up requirements	59
C.6.8 Testbed preservation metadata	61
Annex D : Additional technical background and context	69
D.1 Technical background and analysis	69
D.1.1 Digital recordkeeping and archiving	69
D.1.2 Technical dimensions of the digital preservation problem	70
D.1.2.1 Digital media suffer from physical decay and obsolescence	70
D.1.2.2 Digital records depend on software	71

D.1.2.3 Additional considerations	72
D.1.3 Criteria for an ideal solution to digital preservation	72
D.1.4 Analysis of existing and previously proposed approaches	73
D.1.4.1 Reliance on printing	74
D.1.4.2 Reliance on standards	75
D.1.4.3 Reliance on computer museums	77
D.1.4.4 Reliance on migration	78
D.1.4.5 Reliance on “viewers”	80
D.1.4.6 Reliance on “digital archaeology”	80
D.1.4.7 Reliance on saving the bits.	81
D.1.4.8 Reliance on emulation	82
D.1.4.8.1 Details of the emulation approach	83
D.1.4.8.2 Efficiency of the emulation approach	85
D.1.4.8.3 Natural experiments related to emulation	86
D.1.4.8.4 Unanswered questions about emulation.	88
D.2 Comparing alternative preservation approaches	89
D.3 Scope and limitations of the study	92
D.3.1 Focus on technical aspects of digital archival preservation	92
D.3.2 Focus on mainstream problems that generalize	93

1. Overview

This report presents the results of a short study undertaken by RAND-Europe for the National Archives and Ministry of the Interior of The Netherlands. The primary goal of the study was to define a strategy and framework for the long-term management and preservation of digital governmental records, taking into account policy, organizational, and technical aspects of the problem. The resulting framework includes a preservation process and an infrastructure to support that process, as well as a “testbed” to be developed by the Dutch National Archives as an experimental environment in which specific digital preservation techniques can be prototyped and evaluated on representative types of digital records. The results of this study should be applicable to all government organizations and all types of digital documents, data, and records.¹

The report consists of a body plus four Annexes. The body presents a concise overview of the study’s motivation, scope and recommendations. Annex A presents the recommended digital preservation strategy, Annex B presents the recommended framework, and Annex C presents the proposed testbed (including an initial set of research questions, an experimental design for answering these questions, and specifications for the resources required to implement the testbed). The body of the report and Annexes A, B, and C constitute a full discussion of the results of the study. Additional technical background, context, and further justification of the scope and conclusions of the study are presented in Annex D.

2. Motivation

Digital records are vulnerable to loss in ways that do not apply to traditional paper records, where the “loss” of records is taken to include their becoming physically unreadable or inaccessible, their becoming uninterpretable, or their becoming separated from the contextual information required to make them meaningful. While the vulnerabilities of traditional, paper records are well understood and can to a large extent be successfully mitigated by procedures and techniques developed over the centuries, the new vulnerabilities of digital records are not yet fully appreciated or comprehended—and in many cases defy simple solution.²

¹ Except where relevant, we do not distinguish between documents and data in this report, using the former term to encompass the latter. Also, although the term “electronic records” is often used as a synonym for digital records, the latter more accurately reflects the concerns of this report. Digital records can in principle and practice be represented in non-electronic form (such as on optical or quantum media) whereas electronic documents (such as FAX) may not be digital in any meaningful sense.

² The term “digital longevity” has therefore to date been something of a contradiction in terms. Despite the fact that digital information can be copied perfectly, which should in theory make it eternal, it can quite easily become physically unreadable or uninterpretable meaningless, prompting the ironic observation that “Digital information lasts forever—or five years, whichever comes first.” (Jeff Rothenberg. “Ensuring the Longevity of Digital Documents,” *Scientific American*, Vol. 272, No. 1, January 1995, pp. 42-47.)

Preservation involves more than simply preserving the binary digits (bits) that represent digital records: the authenticity and meaning of the records must also be preserved. That is, we must:

Enable reliable, authentic, meaningful and accessible records to be carried forward through time within and beyond organisational boundaries for as long as they are needed for the multiple purposes they serve.³

Hans Hofman of the Dutch National Archives rephrases this as a question:

How can we carry authentic records through time in a usable and understandable way?⁴

This report offers a strategy for developing ways of preserving digital records that will ensure that they remain authentic, usable and understandable far into the future.

The preservation problem can be viewed in various ways.⁵ One such view is shown in Figure 1: this IDEF0 diagram⁶ depicts the activity or process of preserving digital records.⁷ Reading the diagram from left to right, the preservation process takes records and information about them as its input and produces as its output “preserved” records with appropriate metadata. The diagram illustrates a number of facts about the preservation process: (1) that it takes as input not only records themselves but also their context (including the business processes that produce and use them), (2) that it is controlled and constrained by archival and recordkeeping requirements, (3) that it relies on available preservation technology, and (4) that it produces metadata accompanying the records it preserves.

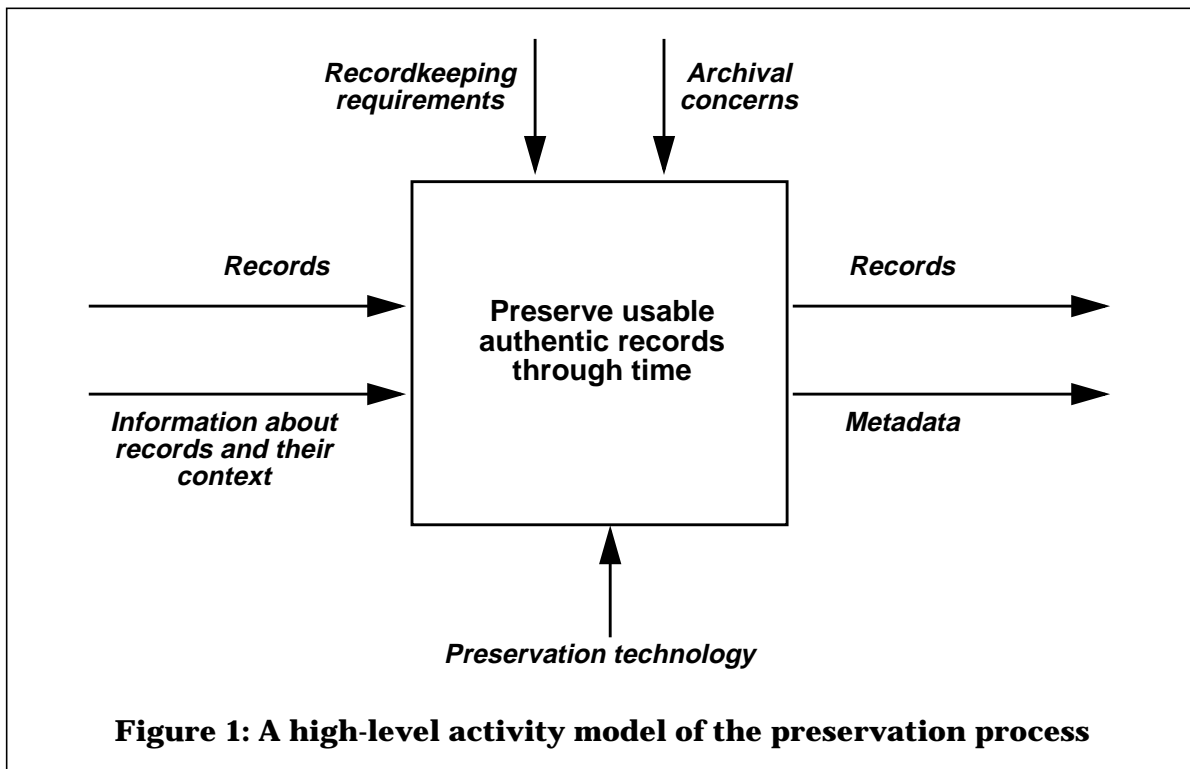
³ Sue McKemmish (Monash University, Australia) and Dagmar Parer (previously of the Australian National Archives) in *Towards Frameworks for Standardizing Recordkeeping Metadata*.

⁴ The term “authenticity” is used in this report to encompass both integrity and authentication, where “integrity” is often used to denote that property of a record that ensures that it has not been changed or corrupted in any meaningful way, and the term “authentication” is used to mean the verification that a record was indeed generated by its purported author or organization at the time and under the conditions of its purported generation. Our broad usage of “authenticity” is closer to the meaning of the word in general parlance: in this report, it denotes the full range of properties that establish the legitimacy of a record. Authenticity issues are discussed in further detail in Annex A, Section A.1.2.

⁵ This report adopts the “records continuum” perspective favored by the Dutch National Archives.

⁶ An IDEF0 diagram consists of boxes that represent activities that are connected to other activities by arrows that represent inputs, controls, outputs, and mechanisms of those activities. Inputs enter an activity from the left and are processed by the activity, typically being transformed into outputs, which exit from the right. Mechanisms enter from below and are used by the activity in doing what it does. Controls enter from above and constrain or direct the activity. For further discussion of IDEF0, see U.S. FIPS 183, *Integration Definition for Function Modeling (IDEF0)*, December, 1993.

⁷ A “preservation process” implements an abstract “preservation function” and may in turn involve the use of a concrete “preservation system” of some kind.



In addition to describing the preservation process as an activity, the diagram can also be interpreted as describing the design of that process. From this perspective, the diagram corresponds to the preservation strategy described in Annex A: the controls entering the activity box from above correspond to the “top-down” flow of that preservation strategy, while the mechanisms entering from below correspond to the “bottom-up” flow of the strategy.

3. Scope of the study

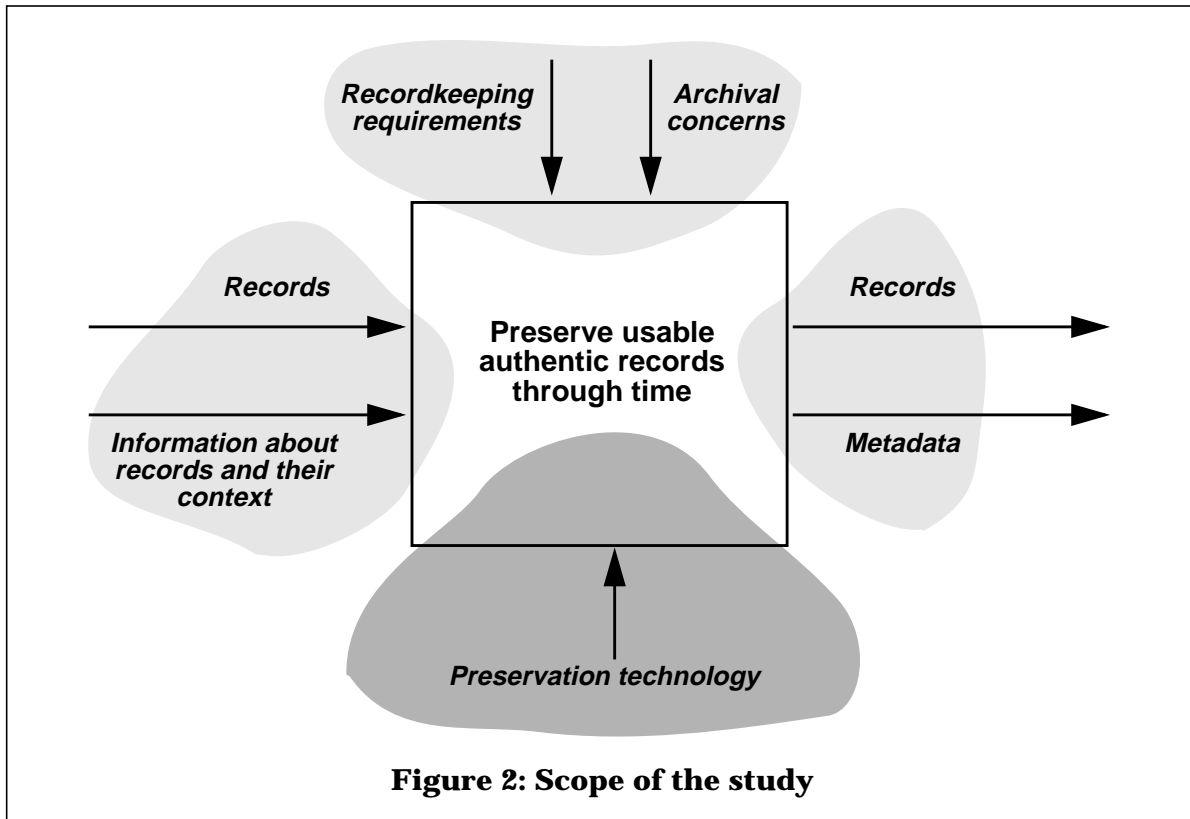
As Figure 1 suggests, the problem at hand has many aspects. Given the short time frame and limited budget of the study, our first task was to identify the issues most relevant to the problem and analyze those that were crucial to performing the study itself, while indicating which others were out of scope.⁸

The scope of the study is suggested by the shaded areas in Figure 2. The primary focus was on the technological issues involved in preserving authentic digital records through time, as indicated by the darker shading in the figure.⁹ Other relevant aspects of the problem (indicated by lighter shading) were taken into account to the

⁸ Further discussion of some of these scope issues can be found in Annex D, Section D.3.

⁹ In particular, we are concerned with records only after they have been explicitly represented as records. Furthermore, we are concerned not with transitory records but with those whose retention schedules require them to be preserved longer than the likely “time-to-obsolence” of their native forms.

extent necessary and feasible. The client directed that the study should focus on the technical aspects of digital preservation, rather than on the intellectual or



organizational issues surrounding preservation or on ancillary recordkeeping issues such as those concerning the maintenance of contextual metadata¹⁰ or other archival concerns such as selection, description, or access. However, many of these issues have a direct bearing on digital preservation, making it impossible to ignore them entirely.¹¹ They are therefore discussed where necessary throughout this report.

Some of the issues that were considered out of scope for the study nevertheless have implications on which the study's assumptions or conclusions rely: in such cases, we identified what we call "reverse requirements" which this study levies against other projects or analytic mechanisms (including the project on the intellectual aspects of recordkeeping that is being conducted by the National Archives and Ministry of the Interior, as well as the testbed project that is proposed as a follow-on to the current

¹⁰ Note that we feel it is important to distinguish between context (i.e., the conceptual environment in which records exist) and metadata, which can be used to represent context. We use the former term to refer to the conceptual collection of facts, relationships, uses, processes, etc. that comprise the recordkeeping and archival environment, whereas we reserve the latter term for concrete representations of information describing that environment.

¹¹ In particular, it is important to understand which attributes of digital documents must be preserved in order for them to function as authentic, meaningful records.

study). These reverse requirements represent contextual questions whose answers were logically required by this study but which the study itself was unable to answer due to its short time frame and limited budget. Where necessary, these reverse requirements were used to “parameterize” the study’s conclusions and recommendations.

4. Summary of results and recommendations

The results of the study consist of three related products plus a number of recommendations and observations and a proposed set of next steps to be undertaken by the National Archives.

4.1 Products of the study

The three major products of the study are: (1) a strategy for preserving digital records, (2) a framework in which to perform such preservation (consisting of a preservation process and a proposed supporting infrastructure and repository), and (3) the design of an experimental testbed in which to try out key ideas and answer key questions related to the preservation of digital records.

The strategy and framework are independent and equally important. The strategy provides a method of deriving requirements for preserving digital records and determining which technological preservation approaches can meet those requirements. The framework provides an overall preservation process for digital records, derives infrastructure and repository requirements for that process, and explains how to use the testbed to perform experiments and build prototypes to make the framework more concrete.

The preservation strategy proceeds both top-down and bottom-up as suggested by Figure 1. From the top, it begins by identifying archival recordkeeping requirements, from which “authenticity criteria” are derived.¹² These criteria are used to determine which attributes of digital records must be preserved; these in turn determine a set of properties that characterize technological approaches that would be capable of preserving the required attributes. Proceeding upward from the bottom, the strategy identifies technological approaches that have the requisite properties and can therefore preserve those attributes required by the given authenticity criteria. Finally, the strategy selects one or more preservation approaches that meet the requirements: other factors being equal, approaches that are less constraining are preferred, since they satisfy the greatest range of authenticity criteria.

The preservation framework includes the design of a generic, “parameterized” preservation process, a supporting infrastructure and repository for this process, and

¹² Article 12 of the 1995 Archival Ordinance for The Netherlands may have the effect of causing government agencies to define implicit authenticity criteria of this kind.

an experimentation process that utilizes the testbed to supply values for the parameters of the preservation process. The testbed includes a set of initial questions to be answered, an experimental design for answering these questions, and an infrastructure to support experimentation.

4.2 Observations and recommendations

The nature of digital records makes the problem of preserving them fundamentally different from that of preserving traditional records. A traditional document is a physical artifact: saving that artifact preserves all aspects of the document that are inherent in its physical being.¹³ In contrast, it is unclear what it means to save a digital document: there is as yet no accepted definition of digital preservation that ensures saving all aspects of a document.¹⁴ Choosing a particular digital preservation method or technology determines which aspects of a document will be preserved and which ones will be sacrificed; this technological choice therefore entails explicit or implicit decisions about what to save. Whereas we can choose to “just save” traditional documents in their entirety, there is no corresponding option for digital documents: any technological choice we make has inescapable implications for what will (and will not) be preserved.¹⁵ In the digital case, we must choose what to lose.

We consider it axiomatic that preservation without access is meaningless. The primary purpose of an archival preservation process is to allow future users to retrieve, access, decode, view, and interpret records as required (either for ongoing use of the records in some business process or for accountability purposes). Any digital preservation approach may constrain such access by failing to preserve certain aspects of records, though saving digital records in their native forms would entail the fewest such losses.

In order to avoid the inadvertent loss of any aspects of digital records that are crucial to ensuring their authenticity, our preservation strategy involves formulating explicit “authenticity criteria” (i.e., criteria for preserving authentic digital records). In general, these criteria are intended to ensure that preserved records will function as required, both for use in some business process and to answer questions of accountability. Different types of records may have somewhat different specific authenticity criteria: for example, authenticity criteria for databases or compound, multimedia records may differ from those for simple textual records. Yet in all cases

¹³ Though an isolated document may not be a record in its own right, this argument applies equally to records, to the extent that they consist of documents and data (including descriptive metadata).

¹⁴ Only in the simplest cases are digital documents direct encodings of traditional documents; more generally, they are executable programs that behave as documents only when interpreted by appropriate software. (An “executable” digital document is one which *has no meaning as a document* until it is executed, i.e., interpreted.)

¹⁵ As noted above, it is not always possible to save traditional documents perfectly or indefinitely either, but traditional documents degrade slowly over time, whereas the losses that arise from the choice of a digital preservation technology occur the moment a digital document is saved.

the intent of these criteria is to ensure that preserved records retain their original behavior, appearance, content, structure, and context, for all relevant intents and purposes.¹⁶

This implies that different technological approaches can potentially preserve the authenticity of digital records, depending on exactly what is meant by authenticity. Different authenticity criteria produce different preservation requirements that can be met by alternative technological approaches. Conversely, some preservation approaches can satisfy a broader range of authenticity requirements than others, which may make them preferable. Nevertheless, it is important to note that merely stating preservation requirements does not ensure that they can be satisfied by available technology: the need for the current study—as well as for the further research we recommend performing in the testbed—arises from the fact that satisfying these requirements for digital records remains problematic.

Of the digital preservation approaches that have been proposed, migration has received by far the most attention. This involves the periodic conversion of records into new forms as older forms become obsolete, utilizing varying degrees and forms of standardization.¹⁷ Yet the appearance and behavior of digital documents (or even their content, structure or context) may depend on the behavior of proprietary software, making it unclear whether migration or conversion will be able to preserve such documents authentically. Further, the rapid introduction of new digital paradigms coupled with the relatively slow development of standards (and compliant software) makes relying on standards for preservation something of an act of faith. Finally, while migration may appear relatively straightforward, it requires unique, specialized treatment of each different document type, as well as individualized application to every document every time obsolescence necessitates migration.

A potentially promising proposed alternative approach is emulation, which offers a number of theoretical advantages. Emulating obsolete computers on future

¹⁶ Note that this extends the criterion for traditional records, for which it is considered sufficient to retain content, structure, and context. For example, a digital record may exhibit various kinds of dynamic or interactive behavior, may include active (possibly dynamic) linkages to other records, and may possess a distinctive “look-and-feel” any of which may—at least in some cases—be important to retain.

¹⁷ The meaning of the term “migration” in this context has evolved in recent years to include both the copying of bit streams to new storage media as old media decay or become obsolete (sometimes referred to as “refreshing” the media) and the conversion of digital documents into new digital formats as the software systems on which they depend become obsolete (see Annex D, Section D.1.2.2 for further discussion of software dependence). So long as storage technology continues to evolve rapidly, media obsolescence will continue to necessitate refreshing media, regardless of how software dependence is addressed. Format migration (that is, converting data and documents into new digital formats so that they can be accessed by new software) has been performed for decades in computer science, where it is generally recognized as being a complex, labor-intensive and high-risk process, which is only now beginning to be applied to the preservation of digital records. As noted in *The Open Archival Information System (OAIS) Reference Model* (CCSDS 650.0-W-5.0) for archival information systems (available at http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html), “Digital migrations are time consuming, costly, and expose [an archival information system] to greatly increased probabilities of information loss.”

computers would allow running the original software that created or viewed a digital record, even after that software becomes obsolete in the future.¹⁸ This would allow saving digital records in their native form and would avoid the problems of software evolution and paradigm shifts. Furthermore, emulation would require no special treatment of different document types and no processing of individual documents at all, other than saving their bit streams.¹⁹ On the other hand, emulation would require saving software and associated documentation as well as descriptions of hardware computing environments to allow generating future emulators for them.

One of the first priorities of the testbed will be to design and conduct trials of migration/conversion and emulation strategies, using an initial selected sample of digital documents as test material.

The use of metadata will play an important role in any digital preservation scheme. General-purpose preservation metadata has been widely discussed in the archival and library communities, but the design of metadata related to the technology of preservation must await further experimentation and prototyping of specific preservation approaches, as in the proposed testbed.

From the perspective of the preservation process, it is irrelevant at what point in their life records enter the process or what agency has responsibility for them when this occurs. However, unduly delaying the entry of digital records into the preservation process may result in increased cost, inefficiency, and risk of loss, since it may become more difficult to begin preserving records the older they become—and ultimately impossible, when they become inaccessible or unreadable.

4.3 Proposed next steps

To begin implementing the recommendations of this report, we propose that the National Archives undertake the following five initial steps (most of which can be performed in parallel):

- 1) Develop a proposal to fund the creation of the testbed and the acquisition of required resources and personnel
- 2) Establish cooperative agreements with several government agencies to facilitate:
 - Sharing digital records for trial use in the testbed
 - Sharing expertise about testbed materials and related issues

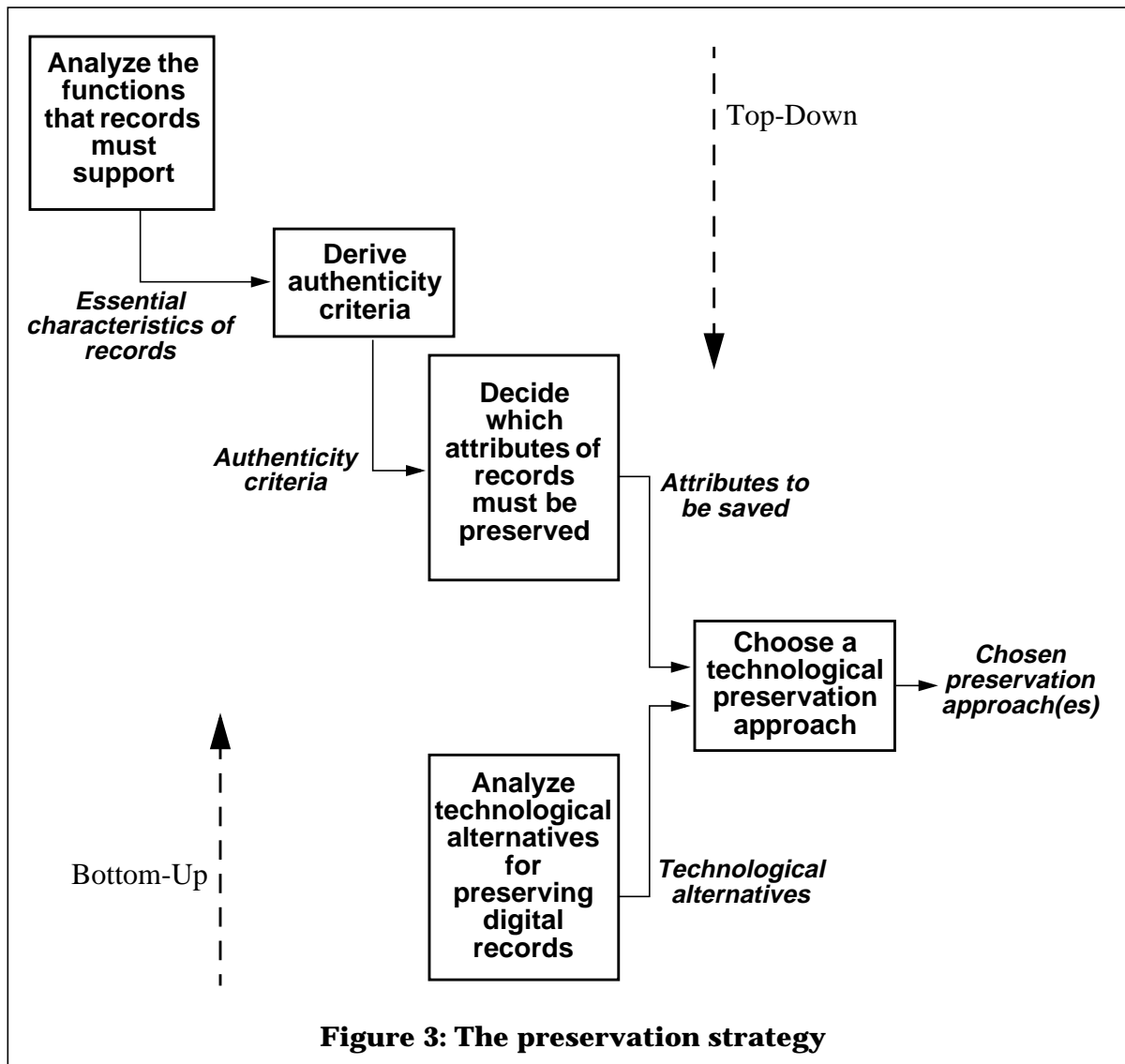
¹⁸ Emulation has also been used for decades in computer science, to extend the effective life of obsolete software and data, though it has yet to be used for the prospective preservation of digital records.

¹⁹ Emulation would not avoid the problems of the decay and obsolescence of storage media, so it would still require “refreshing” media by copying saved bit streams to new media (as would migration).

- Development of evaluation criteria for digital preservation
 - Evaluation of experimental results
- 3) Define a sample of digital records from agencies
 - This sample should be small but reasonably diverse and general
 - Begin to develop “acceptance test” criteria for these document types
 - 4) Budget for, acquire, and develop testbed infrastructure
 - Acquire and install required hardware and software
 - Obtain and allocate personnel with appropriate expertise (within the National Archives, to ensure continuity and build intellectual and technical capital)
 - 5) Perform detailed experimental design of first testbed tasks
 - Designate an “Experiment Architect”
 - Design and begin the first “spiral” of testbed experiments

Annex A: A strategy for preserving digital records

Although many organizations have by now had some experience in preserving digital records, none have done so over the long term, since the oldest such records are relatively young.²⁰ The following offers a strategy for preserving digital records that is based on the intersection of the preservation needs of recordkeeping with the capabilities of digital preservation technology. The strategy proceeds from opposite ends of the problem inward toward a solution: if preservation requirements are thought of as the top and digital preservation technology as the bottom of this space, the strategy can be thought of as proceeding both top-down and bottom-up, as illustrated in Figure 3.



²⁰ Furthermore, the types of digital records for which the most experience has been accrued (i.e., tabular numeric data) tend to be less problematic from a preservation perspective than newer digital forms.

A key aspect of the strategy is that neither the top nor the bottom of this space are taken to be fixed, static points. As argued below, specific preservation requirements are determined by the functions that preserved records must continue to support, with different requirements having different implications for what must be done to ensure the preservation of authentic, meaningful records.²¹ Similarly, the technology underlying digital preservation will continue to change as information science evolves, producing both new constraints and new alternatives. The strategy is therefore intended as a repeatable or ongoing process that may lead to different specific answers in different environments and at different stages in the future. This process (and therefore the strategy itself) can be thought of as depending on a number of “parameters” whose values are context-dependent.

Since the strategy proceeds both top-down and bottom-up, it has two starting points: the functions that digital records must support and the technical alternatives for preserving them.

A.1 The top-down flow of the strategy

The strategy proceeds in a top-down manner to derive requirements for digital preservation technology. The first two steps of the strategy (analyzing the functions that digital records must support and deriving authenticity criteria from this analysis) are presented for logical completeness. In practice the strategy may be initiated by assuming authenticity criteria, thereby eliminating these two steps (i.e., performing them implicitly). Nevertheless, we feel that the inclusion of these steps is important because they provide the logical starting point for the top-down flow of the strategy: without them, this starting point would be arbitrary and difficult to justify.

A.1.1 Analyze the functions that records must support

The top-down starting point for the strategy is the analysis of the business process and associated functions that must be supported by authentically preserved digital records. These functions may extend well beyond the simple retention of readability. They must include functions that support the records’ use in ongoing business processes (whether by the agency that created the records or some other organization) as well as those that support their use for accountability purposes.²² While many of these functions may be similar across all records, it is particularly important to

²¹ For further discussion of what it means to preserve a record, see Annex B, Section B.1.

²² While these two sets of functions may overlap, they need not be identical. The future business processes in which a preserved record may participate implicitly define a set of functions that the preserved record must support, which may not include all of those functions that were required by the processes in which the record has previously participated. Yet future use of the record for accountability purposes may require it to support these earlier functions as well. For example, it might be necessary to ascertain whether a manager who made some decision based on this record failed to consider information that was linked to the record in a way that would have been readily accessible using its original software, even if that information is irrelevant for all future business processes in which the record may participate.

perform this analysis for digital records, since the choice of appropriate preservation technology may depend on the outcome.²³ It is important to make this analysis explicit, even though some agencies may already have performed it implicitly.²⁴

This analysis of the business process functions that digital records must support produces explicit assumptions on which a given invocation of the strategy relies. By monitoring these assumptions, future users of the strategy can decide whether the results of a past invocation may have become invalidated, implying that the strategy should be reinvoked with updated assumptions.

The result of this step should be an explicit set of the essential characteristics of records, i.e., those characteristics that must be preserved in order for the records to retain their required functionality in relation to the business processes they support.²⁵

A.1.2 Derive authenticity criteria

The strategy next derives specific authenticity criteria from the analysis of the functions that must be supported by the digital records to be preserved.²⁶ These authenticity criteria determine which capabilities digital records must retain in order to be considered “authentic” (using that term in its broadest sense).²⁷

As discussed in Section 4.2 of this report, authenticity criteria are intended to ensure that preserved records will function as required, both for use in ongoing business processes and to answer questions of accountability. Different types of records may have different specific authenticity criteria: for example, authenticity criteria for databases or compound, multimedia records may differ from those for simple textual

²³ The archival goal is to preserve everything that is appraised as having archival value, without making restrictive judgments about what future users may need to do with those records. That is, archival preservation should not introduce biases or preempt required future use. Yet as discussed in Section 4.2 of this report, choosing a technological preservation approach may inescapably entail choices about what will and will not be preserved. It is therefore important to make the functions that digital records must support explicit so that the capabilities of alternative preservation approaches can be evaluated against these requirements.

²⁴ In particular, such implicit analysis may be performed when government agencies in The Netherlands comply with the forthcoming requirements arising from Article 12 of the 1995 Archival Ordinance.

²⁵ Note that using this analysis to generate preservation requirements assumes that records can be preserved in whatever ways are required. However, this assumption is no more circular than any other requirement: posing any requirement for a hypothetical process implicitly assumes that it is possible to create a process that can satisfy that requirement.

²⁶ The forthcoming requirements arising from Article 12 of the 1995 Archival Ordinance may cause government agencies to define implicit authenticity criteria, just as it may cause them to implicitly define the functions that digital records must support, as mentioned in note 24 above.

²⁷ Any authenticity criterion is an ideal, which may not be fully achievable under a particular set of technological and pragmatic constraints. Nevertheless, this defines the goal to which any preservation approach must aspire.

records. Yet in all cases the intent of these criteria is to ensure that preserved records retain their original behavior, appearance, content, structure, and context, for all relevant intents and purposes.²⁸

A.1.3 Decide which attributes of records must be preserved

It may be possible to capture many aspects of a record and its context as metadata.²⁹ In fact, metadata may be the *only* way to capture many aspects of the context of a record's generation and original use within the business processes that created or utilized it, since the record itself may contain little or no information about that context. However, certain aspects of digital records (such as their exact form or layout, their precise visual appearance, their interactivity, and any dynamic behavior they may exhibit, such as animation or the execution of embedded program scripts) may be difficult or impossible to capture in metadata.³⁰ The core of the technical problem of preserving digital records is therefore to identify those attributes of such records that must be preserved and to find ways of preserving them, whether through the use of metadata or by other means.

Authenticity criteria determine which attributes of a digital record must be preserved and which ones need not be. The attributes derived by this process must be concrete and specific enough to serve as the evaluative criteria against which alternative technical approaches to preservation will be judged. In some cases, these attributes will represent relatively simple aspects of digital records, such as their text-stream contents or pictorial or tonal resolution. However, in other cases, these attributes may represent more subtle aspects of records, such as their interactivity or their ability to convey particular topological structure relationships among their components or between themselves and other, related records.

The full set of attributes that must be taken into account when invoking the strategy will depend on the types of digital records under consideration. The subset of these

²⁸ As discussed in note 16 of this report, digital records may embody various kinds of behavior that may, at least in some cases, be important to retain. In particular, they may exhibit dynamic or interactive behavior that is an essential aspect of their content, they may include active (possibly dynamic) linkages to other records, and they may possess a distinctive "look-and-feel" that affects their interpretation.

²⁹ It might even be argued that the entire significance of a record (as a record) can be completely captured in metadata, making it unnecessary to save the record itself. This extreme view would transform the problem of preserving digital records into one of capturing their essential attributes as metadata and then preserving the metadata. While preserving metadata might still present some challenges, it would not be nearly as complex a problem as preserving digital records themselves, since the form of the metadata would be entirely under the control of the archivist (or whoever creates it). This would reduce the preservation problem to one of preserving the restricted forms of digital representation used for metadata, making it unnecessary to preserve digital information of arbitrary form. For the purpose of this study, we assume that records *cannot* be entirely represented by metadata and that preserving digital records of arbitrary form will therefore remain necessary, at least in some cases.

³⁰ In such cases, capturing the detailed behavior of a record in metadata may be more costly or difficult than preserving that behavior in the record itself.

attributes that are designated as requiring preservation will in turn depend on the authenticity criteria that are derived from the analysis of the business functions that the chosen types of digital records must support.

A.2 The bottom-up flow of the strategy

The strategy proceeds in a bottom-up manner to derive the capabilities of alternative digital preservation approaches.

A.2.1 Analyze technological alternatives for preserving digital records

At any given time in the evolution of information science, there may be a different set of technically feasible, economically viable technological approaches to digital preservation. Although our strategy is intended for initial use in the present, it is designed to be capable of being reapplied in the future if its initial results prove inadequate to the evolving demands of digital preservation or if technological evolution creates attractive new approaches that appear to offer advantages over whatever preservation approach is chosen initially.

In general, preservation approaches whose technological properties are least constraining are to be preferred, since they satisfy the greatest range of authenticity criteria. What we mean by “least constraining” here is that a preservation approach should ideally preserve as many attributes of the original digital record as possible, thereby making it unnecessary to evaluate in detail which attributes must be preserved.

The current set of available technological alternatives for preserving digital records is analyzed in Annex D, Section D.1.³¹ Of the preservation techniques that have so far been proposed, the one that has received by far the most attention is migration, which implies the periodic conversion of records into new forms as older forms become obsolete, utilizing various degrees and forms of standardization. A potentially promising proposed alternative—which has not yet been developed in full detail—is the use of emulation to allow running original software to access obsolete records. We recommend that these two approaches be considered first in applying this strategy (see Annex C).

Whatever alternative technological approaches to digital preservation are chosen for consideration by a given invocation of the strategy, they should be prototyped and analyzed in sufficient detail to enumerate their relevant properties for use in the final step of the strategy, i.e., mapping those attributes of digital records that are required

³¹ See also Rothenberg, J., *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation: A Report to the Council on Library & Information Resources (CLIR)*, January 1999, which can be read at <http://www.clir.org/pubs/reports/rothenberg/contents.html> or downloaded from <http://www.clir.org/pubs/reports/rothenberg/pub77.pdf>.

to be preserved into the properties of technological approaches that can satisfy those preservation requirements. The testbed is designed to perform this function.

A.2.2 Choose a technological preservation approach

The top-down flow (which begins with analyzing the functions that digital records must support) converges with the bottom-up flow (which begins with technological preservation alternatives) in this final step of the strategy.

The digital record attributes that must be preserved determine a set of properties that characterize those technological approaches that are capable of preserving the required attributes. This mapping of record attributes to technological properties selects a set of candidate technological approaches. In some cases this mapping may be straightforward. For example, the need to preserve the textual content of records implies that a preservation technology must be capable of either preserving the textual encoding of the original along with sufficient information to allow interpreting that encoding in the future or translating that encoding into future encodings in a way that can be guaranteed and proven to result in no loss or corruption of the original content.

In other cases, however, the properties of a preservation technology that are implied by the need to preserve certain record attributes may be more complex and less obvious. For example, preserving the look-and-feel of a digital record or preserving its original authentication (for example, to allow meaningful, authentic comparison of signatures within related records) would levy more elaborate requirements on any technology that claims to preserve these record attributes. Similarly challenging would be the possible need to preserve the original privacy attributes of a digital record, for example, because it may be impossible (until sometime in the future) to decode the record's original privacy scheme in order to re-encode it using the privacy scheme of the recordkeeping system into whose custodianship it has been placed.

Mapping the attributes of records that must be preserved to the properties of preservation approaches requires an understanding of the technological preservation approaches that are available or possible, since the mapping must result in meaningful, relevant properties of those approaches. It is for this reason that we say that this mapping is the convergence of the downward and upward flows of the strategy. The testbed will perform this mapping based on expert evaluation of the results of prototyping.

An analysis of the results of this mapping should produce a set of acceptable technological candidates that are capable of satisfying the relevant preservation requirements. If this analysis yields more than one acceptable alternative, then (as discussed above) approaches whose technological properties are less constraining are generally to be preferred, since they satisfy the greatest range of authenticity criteria. Pragmatic factors (such as difficulty and risk of implementation, projected cost of use,

or administrative complexity) may also be used to help decide which candidate will be used. In some cases, different approaches may be found suited to different subsets of the digital records to be preserved, but a single approach is to be preferred if possible, because of the simplicity and probable economy of scale offered by maintaining only a single preservation technology. At the current time, it is anticipated that choosing among acceptable technological candidates will be far less of a problem than finding even a single acceptable candidate that satisfies the preservation requirements.

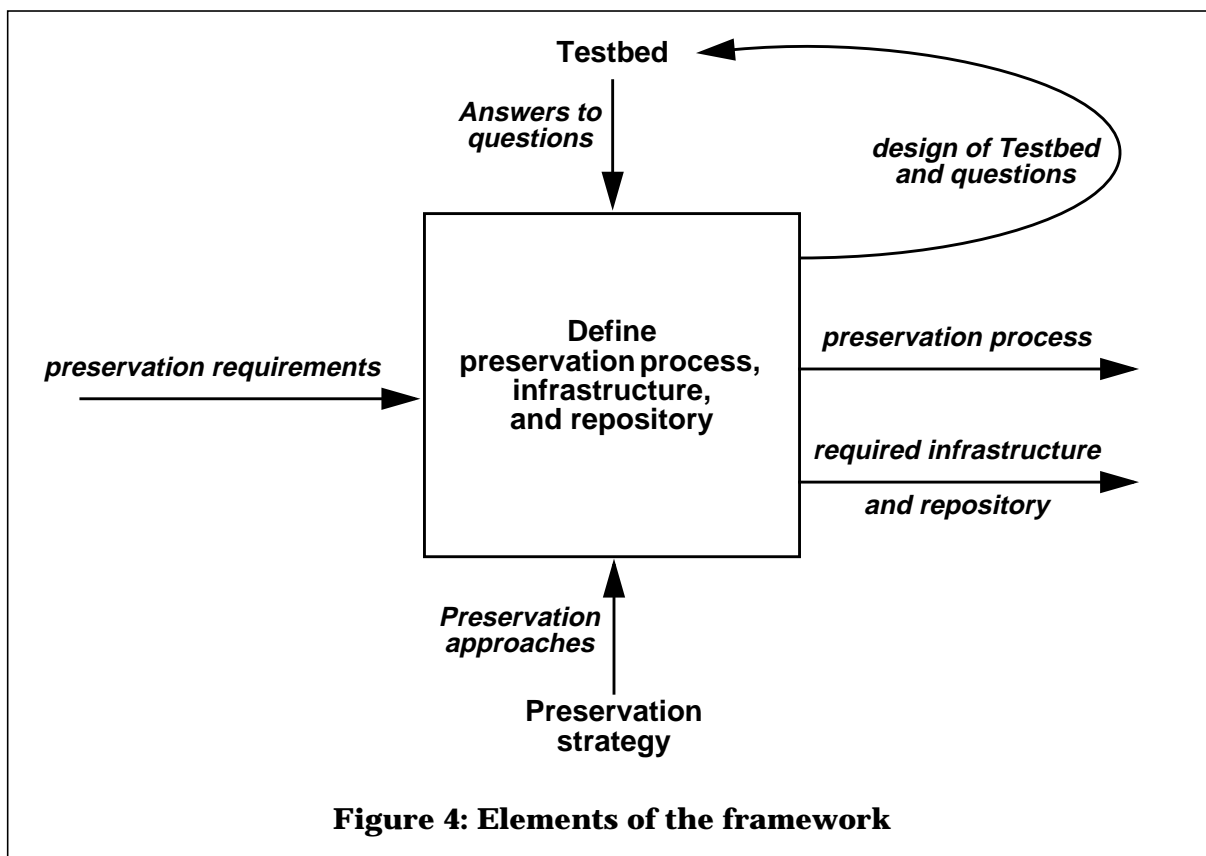
A.3 Integration of digital and non-digital archives

The strategy presented here is intended to address the preservation of digital records. The preservation of non-digital, traditional records is a separate issue. Ultimately, a national preservation program should integrate digital and traditional records. However, the preservation issues associated with these two classes of records are so different that it is questionable whether an integrated preservation strategy could mean much more than integrated management, administration, or funding—and whether even that would be advisable, given the very different methods and constraints of the two types of preservation.

Nevertheless, at least one aspect of preservation should be integrated across traditional and digital records, namely access. An agency or individual seeking records for a certain purpose may not and should not need to know whether the relevant records are digital. Access to all government records—whatever their form, whoever controls and manages them, and wherever they reside—should be uniform and as seamless as possible. This rests largely on the integration of finding aids and other metadata for all kinds of records into a single metadatabase, which should be in online form to facilitate access. The vision of a national Corporate Archival Database for The Netherlands provides just such an integrated metadatabase (where a records continuum perspective suggests interpreting “archival” as applying to records at various stages of their life). Though they extend beyond the scope of this study, the needs of an integrated metadatabase of this kind appear to be quite compatible with the metadata requirements for digital preservation (as discussed further in Annex B, Section B.1.7).

Annex B: A framework for preserving digital records

The preservation strategy presented in Annex A provides a mechanism for identifying and selecting technological approaches that are suitable for preserving digital records. The framework presented here describes the processes and supporting environment in which the approaches chosen by the preservation strategy can be used to perform the ongoing preservation of digital records. The relationships among the elements of the framework are illustrated in Figure 4. The framework employs an experimental environment (referred to as a “testbed”) to answer key questions required to develop a working process for preserving digital records and a repository for such records. The framework includes the design of a generic, “parameterized” preservation process and supporting infrastructure and repository, as well as an experimentation and prototyping process that utilizes the testbed to determine the values of the parameters required to make the generic preservation process and its associated infrastructure and repository concrete and specific. The framework informs the design of the testbed, which informs the implementation of the framework, as shown by the loop in Figure 4.³² The testbed is described in Annex C.



³² The framework articulates assumptions for the testbed, thereby removing these assumptions from the testbed itself, though the testbed may change these assumptions. (Note that this aspect of the figure does not conform to IDEF0, in which an activity's outputs cannot feed back into its controls in this way.)

B.1 A generic, parameterized digital record preservation process

For the purposes of this generic preservation process, the boundary (if any) between day-to-day recordkeeping and archiving is immaterial: the process does not require a notion of archival “transfer” or “acquisition” of records.³³ Digital records may require active preservation while they are still in regular use in the agencies that generate them, often far earlier than the time when traditional records were transferred to archives.³⁴ In addition, whereas transferring traditional records to archives typically involves an unambiguous moment of physical transfer, there need be no such event in the case of digital records. The physical location of a digital record is inherently ambiguous, since multiple identical, equivalent copies of a record may exist, each of which may be distributed, partially replicated, and/or dynamically reallocated among multiple physical locations in such a way as to make the “location” of the record quite meaningless. Furthermore, even when it can be defined, the physical location of a digital record bears no necessary tie to the logical control or custodianship of that record, since digital information can easily be managed remotely.

Neither does the physical or even the logical location of a digital record bear any necessary relationship to the way it can be accessed or used. The transfer of traditional records was generally intended to move those records from the agencies in which they were generated or captured—where they were likely to be accessed on a relatively frequent basis in the course of the official functioning of such agencies—to the archives when they were less likely to be needed by the agencies themselves and more likely to be requested for accountability or historical purposes. The different physical locations and facilities of these institutions may have been better suited to these different purposes, and transfer to the archives allowed records from disparate sources to be found in a single place (as well as allowing preservation techniques to be applied to records whose generating agencies may no longer have had any incentive to preserve them or may themselves no longer exist). Yet the distributed storage, control, and accessibility of digital records blurs such distinctions, logically decoupling issues of access and use from issues of control, management, custodianship, and preservation. Digital records can be managed, controlled, and accessed as desired regardless of where they reside.³⁵ On the other hand, preservation must be considered much earlier in their life than is the case for traditional records: whether preservation is undertaken by the agencies that

³³ Preservation is normally based on the appraised value of a record series, which is derived from the role that the series played in some business process. In general, records should be preserved as soon as they are appraised as having archival value, regardless of when in their lifetime this occurs. This strategy is a natural outgrowth of the records continuum perspective adopted throughout this report.

³⁴ In many cases, if such intervention is not undertaken for digital records, they will become unreadable or inaccessible by the agencies that must use them, let alone by others. This may require agencies to think about the preservation of their digital records, whereas they may have been able to ignore this aspect of their traditional records, leaving it to archivists to worry about preservation after acquiring their records.

³⁵ These and other aspects of custodianship can be determined by logical requirements and need not even necessarily be the same for different classes or collections of digital records.

generate or use records or by archives, responsibility must be taken for ensuring that digital records will remain readable and accessible.

From the point of view of the preservation process, it is irrelevant at what point in their life records enter the process or what agency has responsibility for them when this event occurs: we assume these factors will be determined by technological, administrative and pragmatic considerations. For this reason, the time (or phase) at which records enter the process can simply be considered a parameter of the process. However, preservation represents a conceptual “firewall” across which preserved records must be made immutable. If ongoing business processes continue to use a record after it has been preserved, those processes should work with a read-only or “use-copy” of the record, to ensure that the preservation copy (the official, archival record) is never modified. The longer records are used without being preserved, the greater the chance that they will be transcribed or transformed (if only inadvertently) and thereby corrupted. Furthermore, unduly delaying the entry of digital records into the preservation process may result in increased cost, inefficiency, and risk of loss, since it may become more difficult to begin preserving records the older they become—and ultimately impossible, when they become inaccessible or unreadable.

B.1.1 The creation of a digital record

A digital record may enter the preservation process as soon as the record is created, or it may enter it sometime later, possibly after having been transformed in various ways. In order to put the preservation process in its proper perspective, it is useful to distinguish among several logical states of a digital record both prior and subsequent to its entry into the preservation process. To facilitate this discussion, Figure 5 distinguishes between a record and its representation. The innermost ring in the

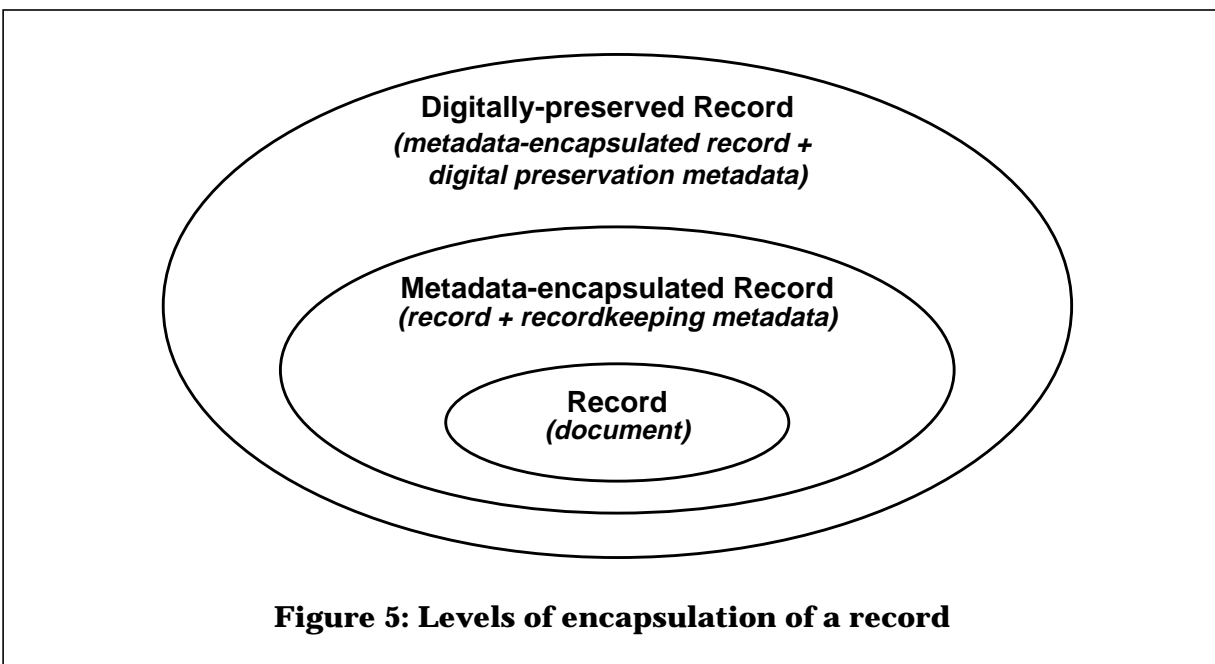


figure corresponds to a record per se, that is, a document or other object that becomes a record by virtue of being generated, sent or received as part of some business process.³⁶ The middle ring corresponds to that record as it is represented in a recordkeeping system, i.e., a “metadata-encapsulated record” having appropriate archival metadata associated with it (including its retention schedule). The outermost ring shows a “digitally-preserved record” which further encapsulates the metadata-encapsulated record with the additional metadata necessary for digital preservation; the focus of this report is on this outermost ring.³⁷ As the figure illustrates, the object to be preserved is the entire metadata-encapsulated record; however, the problematic part of digital preservation involves preserving the digital record itself, i.e., the innermost ring in Figure 5.³⁸

A digital record consists of a document (or other informational artifact of some kind) that is either generated digitally to begin with or is digitized from an initially non-digital form.³⁹ The moment at which a document becomes a record is the moment when it is generated, sent, or received as a record (which may or may not be the moment of its initial creation as a document or the moment when it is “captured”). This is the moment when the innermost ring in Figure 5 comes into being, i.e., when a document becomes a record.

The middle ring of Figure 5 (the metadata-encapsulated record) comes into being when the record is captured by a “recordkeeping system” (or RKS)⁴⁰. The RKS contains metadata representing an archival description of the record and its relevant context, and it associates this information with the record itself.⁴¹ An RKS must consist of at least some physical components: though it need not necessarily use computers or digital methods to perform all of its functions, it must at least provide physical storage and access for those digital records that it contains.⁴²

³⁶ For simplicity, a single record item is shown rather than a record series.

³⁷ Note that preserving transitory records may be fairly straightforward, since their native forms are unlikely to become obsolete within the relatively short period of time over which their retention schedules require them to be kept. This report therefore emphasizes longer-lived records, whose preservation is far more difficult.

³⁸ As discussed in note 29 above, the form of the metadata in a metadata-encapsulated record is under the logical control of the archivist, so it can be chosen to be a form that is easy to preserve, whereas the form of the digital record itself is a given and may therefore prove difficult to preserve.

³⁹ This report focuses on what are often called “born-digital” records, as opposed to “digitized” records, although most records in most organizations at the present time are still digitized.

⁴⁰ Although the term “recordkeeping system” may suggest to some readers a system that is used to store and access records prior to their “transfer” to archival custody, we avoid such notions of “transfer” (and the system boundaries they imply) in this report in favor of the records continuum view.

⁴¹ Although this metadata description logically encapsulates the record, in practice this may mean simply that this description is linked to (or associated with) the record.

⁴² An RKS may be nothing more than a logical entity consisting of information that indicates where digital records reside in the digital systems that store them. In such cases, although the RKS itself would have no physical assets of its own, the physical storage and access devices of the systems that created the records that are logically stored in the RKS must be considered logically part of it.

Note that a record might be produced by (and recorded in) an RKS, in which case the initial system of creation of the record is itself the RKS,⁴³ However, in most cases—at least at the present time—records are likely to be generated in some digital system that is not an RKS, such as a document management system or a general-purpose information system such as a word processor or other “application” system.

The system of creation of a digital record determines its initial digital form. For example, a document may initially have a form determined by the word processing application used to edit it prior to its becoming a record; but that initial form may be no more than a temporary convenience or an artifact of whatever information system the author used to create a draft of the record. For preservation purposes, the relevant form of a record is whatever form it has when it is captured by the RKS: we therefore define this to be the digital record’s “original” or “native” form (from the perspective of the preservation process).

Even after logically entering the RKS, a digital record may be converted or transformed in various ways and/or transferred to other systems by those agencies that generated, captured, or used the record prior to its entering the RKS. If the form of the record is changed by these agencies, its original, native form may be lost, unless it has been preserved by the RKS itself. This motivates the need for a “preservation function” within the RKS, as discussed below.

B.1.2 Entering the digital record preservation process

We say that a digital record “enters” the preservation process, after being appraised as having archival value, when it is formally accepted into what we will call the Preservation System.⁴⁴ We do not define the Preservation System as necessarily being the RKS, though it may be. In fact, we recommend that for efficiency, simplicity, economy of scale, and to minimize risk, digital records should be entered into the Preservation System as early in their life as is feasible, in which case the Preservation System would ideally consist of a subset of RKS functions. Although this is not a logical requirement of the preservation process, we assume throughout this report that this ideal case will be the one implemented, in which case preservation will become one of the functions (the “preservation function”) of the RKS, and the RKS will include the Preservation System.⁴⁵ To avoid confusion, we use the term “RKS” to

⁴³ As noted, from the records continuum perspective, records should enter the RKS as soon as possible after their creation.

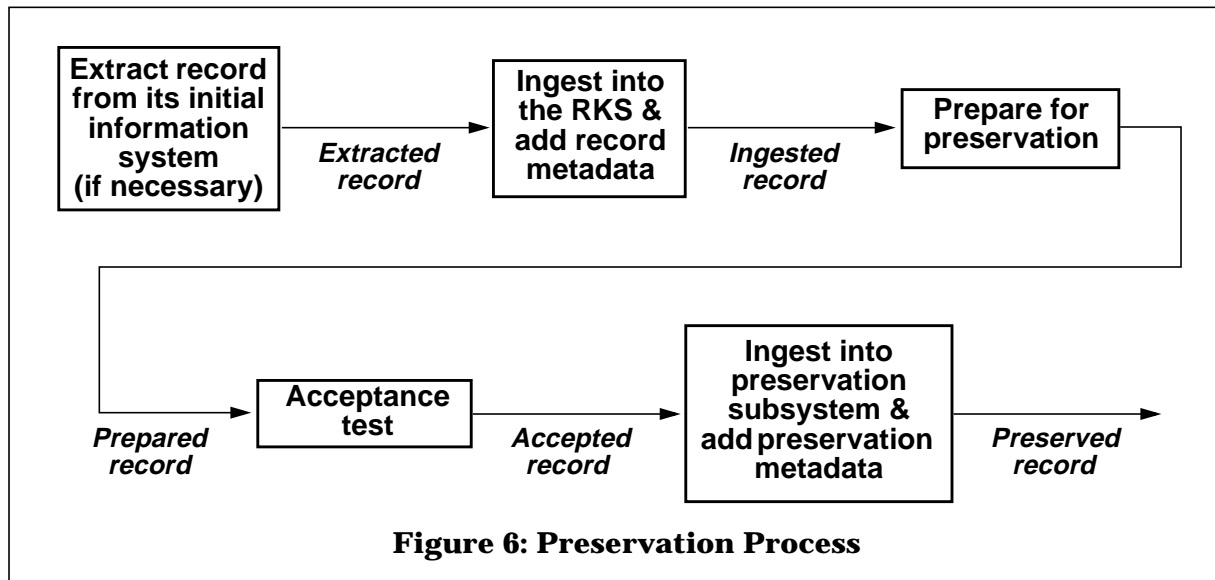
⁴⁴ As noted in Section 2, a “preservation process” implements an abstract “preservation function” which will generally be part of a concrete “preservation system” of some kind. As suggested above, the point in the life of digital records when they enter the preservation process can be thought of as a parameter of the process.

⁴⁵ Note, however, that this will not be the case in the testbed, where a separate, experimental Preservation System will be implemented for records that already reside in an RKS that may not have its own digital preservation function.

denote the full range of functions of the archives and recordkeeping system, reserving the term “Preservation System” to emphasize the preservation function of the RKS.

B.1.3 Extracting a record

In order for a digital record of archival value to enter the Preservation System, several things must happen, as illustrated in Figure 6. First, the record may have to



be “extracted” from the information system in which it resides at the moment when it is entered into the Preservation System. If the information system in which the record resides happens to be the Preservation System itself (that is, the RKS), then this step is unnecessary and no extraction need occur. However, as pointed out above, the record may reside in an application system that was used to create it or in a document management system other than the Preservation System.

Ideally, the process of extracting a record from the information system in which it resides should be straightforward or even trivial, but we identify it as an explicit step in the preservation process because in many cases it may be highly problematic. For example, a record may have been generated by or kept in an information system (possibly even an embedded system⁴⁶) that provides limited retrieval, limited capability for meaningfully accessing records, physical or logical incompatibility with

⁴⁶ For the purposes of this discussion, we consider an “embedded” recordkeeping system to be a recordkeeping system that is embedded within the control system of an automated or semi-automated facility such as a water treatment or power generation plant, a pharmaceuticals production plant, or the Storm Barrier or Air Traffic Control system. Such systems generate and record internal data that constitute records of their operation, i.e., “embedded” records. It is rapidly becoming accepted that such internal data—rather than printed reports or other excerpts or transformations of such data—represent legal records. For example, the U.S. pharmaceuticals industry is increasingly expecting to be required to provide such internal data as records to the Food and Drug Administration (FDA) on demand.

the Preservation System, or has become obsolete to the extent that it can no longer be used at all or that its records can no longer be decoded or understood. This is in fact the situation in many current digital preservation efforts and provides much of the motivation for undertaking the current study. The proposed preservation process cannot solve this problem if the process is invoked so late in the life of a digital record that the information system in which it resides has become hopelessly obsolete. The best way to avoid this problem is to enter digital records into the Preservation System early in their lives—ideally at the moment they become records.⁴⁷

B.1.4 Ingesting a record

Assuming that a record can be extracted from its current information system, it must next be ingested into one of a set of digital preservation forms that are acceptable to the Preservation System. These are forms of digital encoding that are defined for various kinds of digital records, allowing them to be preserved by some preservation approach that is used by the Preservation System. If the record is already in one of these forms, this should require no action, but if it is not, an appropriate form must be chosen, and the record must be ingested into that form.⁴⁸

Any non-trivial transformation or translation of the record itself during ingestion may lose vital information or corrupt the record.⁴⁹ Preservation forms and their attendant preservation approaches must be carefully chosen to ensure preservation of those aspects of records that are deemed relevant and meaningful according to their authenticity criteria. One way to do this is to attempt to preserve digital records in their native forms—for example, for use in an emulation-based preservation approach.⁵⁰ If records are preserved in their native forms, then the set of acceptable preservation forms simply becomes the set of native forms represented by the records to be preserved. For the purposes of defining the preservation process, however, we do not assume any specific preservation approach: instead, we allow the chosen approaches to be parameters of the process. This step of the process therefore describes the general case, in which conversion may have to be performed, though in the ideal case, no such conversion would occur. If a non-ideal approach is chosen,

⁴⁷ Even this may not eliminate the need to extract digital records from the information systems that create them, since the documents or other digital objects that become records may be created using an arbitrary range of application systems. However, if records are entered into the Preservation System as soon as they become records, then it seems unlikely that an application system will have had a chance to become obsolete between the time that a digital document is created using that system and the time that the resulting record is entered into the Preservation System.

⁴⁸ In a sense, ingestion always involves “transformng” a record from its initial form into the required form, but this transformation will be trivial (i.e., the identity transformation) if the initial form is the required form.

⁴⁹ Metadata associated with a record may have to be non-trivially transformed during ingestion, but this should not pose as great a threat, as argued in note 29 above.

⁵⁰ As described in Annex D, Section D.1 and Rothenberg, 1999, op cit, note 31.

specific preservation forms appropriate to that approach must be tested, evaluated, and identified to the recordkeeping community as acceptable.

In order for the RKS to fulfill the special roles of the Preservation System, records should enter the preservation function of the RKS at—or very soon after—the moment they are captured by (ingested into) the RKS. The moment of capture of a record is the moment when the middle ring of Figure 5 (the metadata-encapsulated record) is created: any required descriptive or contextual metadata that is not already provided with the record must be generated by the RKS at this time.⁵¹

The moment when a record enters the preservation function of the RKS is the moment when the outermost ring of Figure 5 (the digitally-preserved record) is created: any required preservation-specific metadata that is not already provided with the record must be generated by the RKS at this time.⁵² As emphasized above, the records continuum perspective argues that records should ideally enter the RKS as soon as possible. Therefore the moment at which a record becomes a record, the moment at which it is captured by the RKS (thereby becoming a metadata-encapsulated record), and the moment at which it enters the Preservation System (i.e., the preservation function of the RKS, thereby becoming a digitally-preserved record) should all ideally be one and the same moment.

B.1.5 Preparing a record for preservation

Whatever the preservation form of a digital record, and whether or not the record must be converted into some such form on entry into the Preservation System, the record must be prepared for preservation. This may entail various steps and varying degrees of effort, depending on the chosen preservation approach for the given type of record, but it will at a minimum involve performing a logical completeness and validation check on the ingested record and preparing technical preservation metadata for the record.⁵³ The logical check must verify the integrity and completeness of the record's content, metadata, etc. That is, it must ensure that all necessary components of the record and appropriate metadata have been accurately captured and that these components—including any linkages among them—are valid.

⁵¹ If some information system other than the RKS initially generated the record—or a document management system managed it prior to its entry into the RKS—these may have produced some of the necessary descriptive and contextual metadata. However, the RKS will generally need to add archival metadata (e.g., appraisal criteria) to make the record into a metadata-encapsulated record.

⁵² As is true for the descriptive metadata discussed in note 51, some of this preservation-specific metadata may also come from previous information systems that generated or managed the record.

⁵³ Since appraisal is normally performed on a record series rather than on an individual record, preparation will normally apply to the entire collection of records in a series. Metadata issues surrounding record series are discussed in Section B.1.7.

The specific technical preservation metadata that must be created for a record at this point will depend on the specific preservation approach that is to be used. Even if records are retained in their native forms or if a record is already in an appropriate preservation form when it enters the Preservation System (in both of which cases, conversion is unnecessary), relevant technical preservation metadata must still be created for the record.⁵⁴

B.1.6 Acceptance testing of records

The prepared record should next be subjected to an “acceptance test” that has been defined for the chosen preservation form. This test amounts to performing a trial preservation/access cycle to verify that the record and its attendant technical preservation metadata are complete and are appropriately prepared for preservation by means of the chosen preservation approach. To perform this test, the chosen preservation approach is applied to the record as it would be in the future, to see if the resulting preserved record remains accessible as it should in the future and whether accessing it in this way retains its key attributes, as defined by its authenticity criteria.

Each pair of preservation form and preservation approach that is accepted by the Preservation System should define an acceptance test of this kind. Each record that is a candidate for acceptance into the Preservation System should pass this test before it is deemed suitable for preservation.⁵⁵ Acceptance testing is performed on the ingested, prepared record, since that is the form in which it is to be preserved; in particular, these tests can be applied only after technical metadata have been created for the record.⁵⁶

Acceptance tests and their attendant verification procedures must be developed as part of the overall preservation process. They may be automated to some extent or they may require human evaluation, though in order to be feasible, the amount of human evaluation must be minimized. If their processing and human requirements are modest enough, these tests should be applied to all records entering the

⁵⁴ At least some of the required preservation metadata will be specific to the preservation approach adopted. For example, if emulation is used, the metadata must include the software needed to view each record (or information about where to find that software), emulator specifications for the required hardware computing environment (or information about where to find such specifications), and documentation that explains how to use the emulation scheme to retrieve the preserved record.

⁵⁵ Failing this test indicates that the record is not properly prepared for preservation by means of the chosen approach (i.e., that applying that approach to the record will not properly preserve it). This should result in some remedial action, such as performing additional preparation, converting the record into a different preservation form, choosing a different preservation approach, or obtaining additional information about the record, as necessary. The term “acceptance test” is meant to imply that if a record cannot pass this test, that record should not be accepted by the Preservation System, since the available preservation approaches are inadequate to preserve it.

⁵⁶ These tests should be performed before expending any effort on creating the additional metadata described below, since such effort would be wasted if the record fails its acceptance test and is rejected.

Preservation System, though in practice it may be necessary to settle for spot-checking of individual records of a given type, in which case appropriate techniques should be employed.

B.1.7 Metadata

Digital records in the Preservation System must be accompanied by appropriate metadata.⁵⁷ Our concern (for purposes of digital preservation) is limited to the technical preservation metadata required by the preservation approaches that are to be applied to records plus the minimum metadata necessary to control and access those preserved records. Additional required metadata (as amply discussed in the literature) will be defined by aspects of the RKS other than its preservation function and are outside the scope of this report.⁵⁸

In any case, relationships between the record item in question and other records may be of crucial importance, whether or not those other records are themselves digital. This makes it essential that preserved digital records be capable of maintaining appropriate linkages or references to non-digital records, which may—at least for the immediate future—reside in traditional archival management systems that are not identical to the RKS posited here. Such linkages must be represented by appropriate metadata in the Preservation System to avoid isolating preserved digital records from non-digital records to which they are logically or contextually connected. Linkage metadata of this kind may not strictly speaking be “preservation metadata” but may nonetheless be required, in order to integrate digital preservation into the larger archival arena.

In some cases, metadata associated with a record by an information system in which the record may have resided prior to its ingestion into the RKS (e.g., an application system or document management system in which the record was initially created or managed) may be deemed irrelevant by the RKS and may therefore be discarded.⁵⁹ Conversely, any additional management metadata specific to the Preservation System must be created at this time.⁶⁰ Similarly, other differences between the

⁵⁷ A digital record can be thought of as existing on at least three levels: the conceptual or intellectual level represents the record as an abstract entity, independent of its embodiment; the logical level represents the record as a related set of components (such as sections, annexes, pages, diagrams, images, logical enclosures or attachments, etc.); the physical level represents the actual digital objects that embody each of the components of the record. Metadata may be required for each of these three levels, in order to characterize the object from a conceptual point of view, explain the logical structure of the record, and map that logical structure to a set of physical digital objects, each of which must itself be described so that it can be accessed, rendered, and combined with other components to produce the record itself.

⁵⁸ Metadata other than preservation metadata can be considered part of the metadata-encapsulated record (i.e., the middle ring of Figure 5) which is preserved as an integral object within the digitally-preserved record (the outermost ring of Figure 5). However, this layered view is merely a logical perspective: all metadata must remain visible at all layers, without being hidden inside inner layers of encapsulation.

⁵⁹ For example, cost or scheduling information specific to the original production environment of a record may be irrelevant to the RKS.

Preservation System and the record's previous information system may necessitate adding or modifying metadata of various kinds on entry into the Preservation System.⁶¹

For preservation purposes, technical preservation metadata must be created and saved in the Preservation System for each record that enters it. Since records are normally preserved in series, it may be necessary to include preservation metadata describing the technical characteristics of the series as well, particularly for records whose technical interpretation may require an understanding of the technical attributes of the series (and system) in which they are embedded. If different records and different types of records are handled uniformly by the Preservation System (which is desirable for reasons of economy of scale as well), most of the technical preservation metadata describing those records will be similar or identical.

The precise metadata requirements of the Preservation System will be determined by the specific preservation forms and technological preservation approaches it uses, which are considered parameters to the preservation process; since these forms and approaches may evolve over time, the preservation metadata in the Preservation System may have to be modified or extended as appropriate. An extensible metadatabase or metadata repository should be designed for this purpose, embodying the best aspects of the archival metadata standards currently being discussed in the archival community.⁶² Interactions and interdependencies may be unavoidable between the technical preservation metadata required for a given preservation approach and management, contextual, or other metadata for the records it preserves, but to whatever extent possible, such dependencies should be avoided in the interest of cleanly separating different types of metadata and allowing

⁶⁰ This includes any metadata required to establish relationships to pre-existing archival management systems.

⁶¹ Even after records are logically ingested into the Preservation System, they may (at least initially) physically reside in recordkeeping systems belonging to the agencies that create or manage them. However, the metadata requirements of the National Archives for records in the archival RKS may differ from those of recordkeeping systems whose scope is restricted to the agencies that create or use those records. For example, the Archives may add explicit metadata representing linkages to related agencies or descriptive information that may be implicit or unnecessary outside the archival context. Archival metadata may therefore not be a simple "view" into pre-existing metadata (where "view" is used in the database sense of selecting, projecting and/or joining existing data).

⁶² These include the development of the University of Pittsburgh's metadata specifications derived from the functional requirements of their Reference Model for Business Acceptable Communications (R. J. Cox, "Re-Discovering the Archival Mission: The Recordkeeping Functional Requirements Project at the University of Pittsburgh, A Progress Report," *Archives and Museum Informatics* 8, no. 4 (1994): 279-300; R. J. Cox, "The Record in the Information Age: A Progress Report on Reflection and Research," *Records & Retrieval Report* 12, no. 1 (January 1996): 1-16; and <http://www.sis.pitt.edu/~nhprc>); the Dublin Core metadata set (see http://purl.org/metadata/dublin_core and H. Thiele, "The Dublin Core and Warwick Framework—A Review of the Literature, March 1995-September 1997," *D-Lib Magazine*, January 1998, <http://www.dlib.org/dlib/january98/01thiele.html>); the Australian Government Locator Service (AGLS) metadata set (1998-07-27, <http://www.ogit.gov.au/aglsindex.html>); and *The Open Archival Information System (OAIS) Reference Model (CCSDS 650.0-W-5.0)* for archival information systems, op cit, note 17.

preservation approaches to be changed without affecting any metadata not directly connected with the technology of preservation.⁶³

An initial set of preservation metadata suitable for the testbed is presented in Annex C, Section C.6.8.

B.1.8 Providing access to preserved digital records

We consider preservation without accessibility to be meaningless. The primary purpose of the preservation process is to preserve understandable and usable digital records. The means of accomplishing this will depend on the technical preservation approach used. In general, however, the packaging of records and metadata in the Preservation System must be sufficient to allow future users to retrieve, access, decode, view, process, and interpret the records as anticipated.⁶⁴

Access consists of finding and retrieving a record (which we consider to be largely out of scope of the present study) followed by “rendering” the record in some way.⁶⁵ The rendered form of a record will in general be determined by the intersection of the demands of whoever is accessing the record with the capabilities of the preservation form of the record and of the Preservation System itself. For some purposes, an accessor may want something as close as possible to the native form of a record, whereas in other cases, a future form that is quite different from the original may be more convenient or useful. This distinction is analogous to that between reading a microfiche reproduction of an ancient text versus reading a modern transcription of that text that uses modern typography, layout, spelling, and language: for some purposes, those aspects of the original may be vital, whereas for record preservation

⁶³ However, the logical relationships between technical preservation metadata and other types of metadata must be made explicit. For further discussion of this issue, see Annex C, Section C.6.8 (including note 111).

⁶⁴ We qualify this statement by saying “as anticipated” because unanticipated kinds of access may or may not be satisfiable by whatever preservation approaches are chosen for the preservation process. The preservation strategy from which this process derives is based on analyzing the functions that digital records must support, which predicts (more or less accurately) what kinds of access to digital records will be required. The technological preservation approach or approaches chosen by the strategy may (though they need not necessarily) limit the kinds of access they enable to those that are required. As we have argued above, preservation approaches that are less limiting in this regard are to be preferred precisely for this reason: since any analysis may be imperfect in predicting the kinds of access that will be required, a preservation approach that is less limiting provides better insurance against surprises. Saving digital records in their native forms would be the least limiting preservation approach, but even this cannot guarantee perfectly preserving all attributes of a digital record. In general, whatever preservation approach is chosen for a given digital record, it is always possible that its access capabilities will be exceeded by unanticipated future access demands. The recommended strategy is therefore to choose a preservation approach that entails as few limitations as possible in order to have the best chance of meeting such unanticipated demands for access—but to attempt to guarantee to satisfy only those access requirements that have been derived from the analysis of the functions that digital records must support.

⁶⁵ We use the term “rendering” as a generalization of “displaying” which may be too limiting to encompass the full range of behaviors of which digital records may be capable.

purposes they may be irrelevant.⁶⁶ As its default behavior, however, the Preservation System should render a record in its preservation form (whether that is the record's native form or some derived form) since this is by definition the most authentic form of the record that the system can provide.

B.1.9 Permanence of the preservation process itself

When a digital record enters the preservation process, it is considered to do so permanently. That is, records should never leave the process once they have entered it. By definition, the preservation process is that which provides ongoing longevity for records. Although the process itself (or the Preservation System that implements it) may evolve over time and may involve various transformations of the records it preserves, such changes in the process or its processing of records are not considered to be cases of records leaving the process or moving into a different process. Records may move into new versions of the Preservation System, which continue to implement the preservation process, but any case of a record leaving the archival preservation process itself after it has entered it is considered an anomaly.

The moment when a digital record enters into the Preservation System is the moment when the initial preservation form of the record is determined.⁶⁷ At any given time in the future, the Preservation System (or its successor) maintains the digital record in some preservation form, which may or may not be its initial preservation form.

Ongoing technical preservation consists of the following steps.⁶⁸

- Ensure that the media on which digital records are stored remain readable (refresh or transfer to new media as needed);
- Evolve new preservation forms if and as needed, along with new validation techniques for those forms;

⁶⁶ Ultimately, the Preservation System might offer the option of transcribing a record from its preservation form into a "use-copy" or surrogate, having whatever form is the most convenient and useful for a given user at a given time in the future. This might be some future, easily understandable form into which the record would be translated, just as ancient texts are translated into the "vernacular" for modern readers. This process may be thought of as performing a "vernacular transcription" or "vernacular extraction" of the record from its preservation form. Additional options might even offer multiple, alternative use-copy forms, including forms from periods intermediate between the (future) present and the original time of entry of a record into the Preservation System. Further discussion of such surrogate forms, however, is outside the scope of this report.

⁶⁷ As pointed out above, this may or may not be the native form of the record.

⁶⁸ As emphasized above, preservation should ideally minimize the conversion or translation of records to avoid their corruption; nevertheless, the steps shown here focus on the places where such conversion may have to be performed if it cannot be avoided. If conversion *can* be avoided, all but the first two steps disappear (the second step may still be needed since future records may require new preservation forms).

- Perform any necessary generic conversion or transformation of records from one preservation form into another if and as these forms evolve;
- Perform any record-specific conversion or transformation of individual records that may be required to maintain their authenticity and accessibility through time;
- Perform any conversion or transformation that may be necessary to keep metadata understandable and usable;
- Revalidate any new preservation forms, converted records, and metadata using an extensible, standardized validation suite;
- Revalidate any converted or transformed records to ensure that they still conform to preservation requirements;
- Create new metadata to document any and all transformations performed.

B.2 Supporting infrastructure and repository

The infrastructure required to support the generic preservation process described above includes the usual elements of any information system: processing power, storage, communications facilities, software, personnel, and administrative arrangements and procedures, all of which in turn rely on that most crucial of infrastructure elements, funding. Part of this infrastructure will be a repository for storing and preserving digital records. The preservation process provides the structure in which these elements are combined to produce the desired goal of preserving and providing access to digital records.

Since the preservation process is as yet specified only generically, with its further elaboration depending on a number of parameters, the design of the infrastructure is similarly parameterized. The testbed described in Annex C is intended to perform experiments to answer research questions that can supply values for at least some of these parameters (others will have to be supplied from exogenous sources).⁶⁹ Similarly, the concrete design of the repository is expected to emerge as one of the results of the testbed.

⁶⁹ In order to conduct its experiments, the tested itself requires an infrastructure, which is logically distinct from (though in some ways similar to) the infrastructure required by the preservation process. To avoid confusion, the discussion here addresses only the infrastructure for the preservation process itself; the infrastructure to support the testbed is described in Annex C.

B.2.1 Processing

The preservation process requires computer processing for a wide range of tasks including: digitizing non-digital records that are to be stored digitally; “ingesting” digital records into the Preservation System; performing validation or “acceptance tests” of records entering the system to ensure that they are initially accessible and can be preserved in an authentic and accessible way; testing, verifying, refreshing, and copying media while performing any necessary coding or storage format conversion; converting and transforming records into new preservation forms when necessary (depending on the preservation approach chosen) and subsequently revalidating the accessibility of those records; generating and converting metadata as needed; processing search queries; and searching for, retrieving, accessing, decoding and “rendering” records with their original functionality intact.

The amount and kind of processing required for refreshing media will be highly dependent on the types of storage media used and the frequency with which they must be tested, verified, copied, and changed. The processing cost of performing these functions may be a significant part of the overall cost of using a particular storage medium, so these costs should be estimated and considered when choosing such media for preservation purposes.⁷⁰

Similarly, the amount and kind of processing required both for the ongoing preservation of records and for accessing and rendering them will be highly dependent on the technological preservation approach chosen. For example, an approach based on migration will require significant processing at ingestion (when records will have to be converted into an acceptable preservation form), at each migration cycle (when records will have to be converted into a new form or redesigned to fit a new paradigm), and for rendering in anything resembling their original form (when records may have to be reconverted back into something similar to their earlier form). By contrast, an approach based on the use of emulation should require minimal processing at ingestion (since records will be ingested in their native form) and no conversion or migration processing, but it will require processing to generate emulators on new computing platforms and to perform emulation during rendering.

The processing power required will also depend on the number and types of records to be stored, but the relationship between the number of records to be preserved and the processing required may be different for different preservation approaches. For example, migration requires processing of all records at every migration cycle, whether they are accessed or not, whereas emulation requires processing only for those records that are accessed, when they are first accessed on a new platform.

The degree to which the required processing must be distributed versus centralized will be an outgrowth of the access usage patterns that emerge for digital records and

⁷⁰ Storage media are currently evolving so rapidly that all decisions involving or depending on them should be reevaluated on a yearly basis and revised every two to three years for the foreseeable future.

of the ways in which the preservation process itself and the digital repository it creates are distributed among the agencies and organizations involved. Since both of these factors are difficult to predict, this aspect of the infrastructure must be allowed to evolve as necessary.

The processing needs of the testbed are described in Annex C.

B.2.2 Storage (repository)

The preservation process requires a repository that can store digital records and their metadata in the required forms⁷¹. The total volume of metadata required for certain kinds of records (such as scientific databases whose contents may require extensive explanation) may in some cases be one or even two orders of magnitude greater than the volume of the datasets themselves. In addition, metadata requirements will depend on the preservation approach or approaches chosen. For example, some approaches may require more explanatory documentation and ancillary material than others.⁷² Similarly, some approaches may require more read/write storage, while others may rely more on immutable storage.

A number of design choices also have significant implications for the amount and types of storage required, such as the degrees of redundancy, replication, reliability, and distribution desired. The desired degree of overall risk-aversion also impacts storage requirements in a number of ways. For example, software or metadata that are logically required to “adhere” to individual records may be replicated with each record or stored once and pointed to from each record. Detailed estimates of the amount and type of storage needed for the repository must therefore be deferred until questions such as these have been answered, whether by the experiments performed with the testbed or from exogenous sources.

The digital repository will consist of a physical store for records plus a metadatabase, which may be a separate entity but must be integrated with record storage to ensure access. Various metadatabase systems are available on the market, but the choice of a specific metadatabase architecture should await the results of experimentation and prototyping to be performed using the testbed. As discussed in Section B.2, the design of the repository is expected to emerge as one of the results of the testbed.

The storage needs of the testbed are described in Annex C.

⁷¹ See the discussion of metadata in Section B.1.7 above.

⁷² For example, emulation requires the storage of original software, documentation for that software, hardware emulation specifications, emulators generated from these specifications, and explanations of how to use the approach to render saved digital records in their original, native forms; however, this applies to all digital records of a given type and requires little or no specific metadata for each record.

B.2.3 Communications

The communications requirements of the preservation process will depend on the specific architecture of the Preservation System, as determined by design decisions about the desired degree of distribution of access, control, management, and administration of digital record preservation, as well as by implementation choices such as physical distribution for reliability or performance reasons. Issues of synchronization may also affect the need for communications if, for example, distributed, replicated repositories will need to synchronize their contents under some particular timing constraint (e.g., to ensure that different users requesting related information obtain consistent results).

Note that these physical architectural aspects of the Preservation System are independent of the logical architecture of the preservation process. For example, the records continuum perspective adopted in this report allows a range of choices about when digital records enter the preservation process and consequently whether ongoing access to those records by agencies will require retrieval from the Preservation System. Yet such choices say nothing about where the records must physically reside: any given logical architecture can be mapped into widely varying physical architectures, each potentially having quite different distribution properties (and thereby entailing quite different communications requirements).

It is possible that large-scale remote, real-time ingest and access may create unexpectedly high demands for communication in the preservation process, and it may be useful to attempt a quantitative estimate of such demands, either in the testbed or in a separate effort. However, in the absence of any surprising results from such investigations, it is expected that existing and projected communications and networking infrastructure in The Netherlands is likely to be adequate for the preservation process.

The communications needs of the testbed are described in Annex C.

B.2.4 Software

The preservation process will require software to provide a repository for digital records and metadata as well as software to perform ingest, acceptance testing, and validation, search and retrieval, “viewers” for various preservation forms, and specialized software to implement the preservation approach or approaches chosen, including any required conversion, access and rendering. Two of the most significant parameters for the preservation process are the sets of preservation forms and preservation approaches to be supported, which account for much of the specialized software required. Prior to specifying these parameters, it is impossible to fully characterize the software needs of the process, except to say that they may be fairly specialized and intensive. For example, if a preservation approach such as migration is chosen that requires ongoing conversion tailored to specific preservation forms, this

is likely to create an intensive, ongoing demand for special-purpose (and therefore potentially expensive) software. Depending on the preservation approach (or approaches) chosen, the preservation process is likely to require a significant amount of specialized software for such preservation-specific purposes as ingest, acceptance testing, and validation. On the other hand, much of the required software, such as repository and search and retrieval software, can probably be obtained in the commercial market.

The software needs of the testbed are described in Annex C.

B.2.5 Personnel

The detailed design of the preservation process and the Preservation System will require a system architect who can work closely with archivists, recordkeepers, and record users in relevant government agencies. The implementation of the system will require close coordination between this system architect and the programmers, contractor, or contractors chosen to build the system. Once in place, the Preservation System will require systems and data administrators who understand archival and recordkeeping issues as well as system issues. The preservation process will require ongoing administrative support by personnel trained in archival and recordkeeping science and attuned to whatever perspective the process ultimately adopts within the records continuum view.

The personnel needs of the testbed are described in Annex C.

B.2.6 Administration

The infrastructure of the preservation process must include administrative arrangements, structure, and support for the management and control of records wherever they logically reside within the records continuum. The continuum perspective allows a wide range of choices for where responsibilities are to be placed for the various aspects of records preservation. Since the choice of a logical position within the continuum is considered exogenous to the current study, we do not suggest a specific administrative structure, pointing out only that whatever structure is established should be consistent with the logical architecture of the preservation process. A flexible mechanism should be created for establishing administrative agreements among the National Archives and the Ministries and other agencies and organizations that generate, capture and utilize records, so that these agreements can adapt as the overall architecture of the preservation process evolves and/or shifts its logical position within the records continuum.

The administrative needs of the testbed are described in Annex C.

B.3 Experimentation and prototyping process

This process is used to determine the values of at least some of the parameters required to make the generic preservation process and its associated infrastructure and repository concrete and specific (others of these parameters are outside the scope of this effort and must be answered by exogenous means). Utilizing an experimental testbed, this process attempts to try out key ideas and answer key questions for the preservation process. The testbed will initially focus on questions whose answers can help specify the preservation process and make the design of its infrastructure and repository more concrete; but the testbed is ultimately expected to help answer a far broader set of questions. The following discussion gives some illustrative examples of the kinds of questions the testbed may initially or ultimately answer, whereas the detailed description of the testbed in Annex C discusses the specific research questions it will pose initially, as well as the experimental approach it will use to attempt to answer these questions.

The testbed will compare and evaluate specific preservation approaches and will generate rich data for answering key questions about the viability of the overall preservation process, such as: Can we cost-effectively preserve digital records by means of the proposed process, enabling the reconstruction of original business processes and the roles the records played in those processes?⁷³ Do such preserved records accurately retain their original relationships to other records? How closely must preserved records recreate their original look-and-feel (and other behavior) in order to fulfill these functions?

Further examples of the kinds of research questions that may ultimately be answered using the testbed include: What types of digital records must be handled in order to produce credible, generalizable results? What kinds of access are most likely to be required of these kinds of records? What are the needs of access and rendering? Which aspects of digital records must be preserved to retain their authenticity? What kinds of metadata filtering must be done when these records enter the preservation system? What kinds of preservation or other metadata must be added to these records? What kinds of contextual models of business function, organizational activity, or actor behavior must accompany these records to make them understandable? How can such models best be represented in metadata? Which technical preservation approach or approaches are viable and how do they compare? What constitutes validation of a preservation approach? What constitutes an “acceptance test” for a given type of digital record? What are the properties, characteristics and relative cost-effectiveness of various preservation approaches? Which preservation approach or approaches should be used? Once a preservation approach is chosen, what specific infrastructure requirements does it imply? What alternative models of archival “transfer” make the most sense? How can Ministries best work with the National Archives to preserve digital records?

⁷³ For an informal analysis of cost issues surrounding specific preservation approaches, see also Annex D, Section D.2.

Note that the testbed will initially focus on key technical questions rather than procedural issues, since the former are considered the most crucial for elaborating a technically viable preservation approach.

Annex C: Testbed

C.1 Rationale

The testbed is envisioned as a facility that combines records and archives management knowledge with technical expertise and technology resources to create the capability for doing significant prototyping and demonstration research addressed to questions about the durability of digital documents.

National and provincial governments and agencies in The Netherlands are major users of contemporary information technologies, collectors and maintainors of large data sets, and providers of critical information or information-based services to citizens, businesses and other customers.⁷⁴ The introduction of networked digital media into the work of such organizations is generally guided by the dual aims of improving the efficiency of mission-based processes while increasing the transparency and quality of those processes for their customers. As a corollary of this transition, an increasing proportion of the official interactions within government agencies or between them and other parties generate records that are digital in nature.⁷⁵

However, because this transition has been oriented around supporting and augmenting primary business processes, much more attention has been focused on the early phases of records generation and use than on later stages.⁷⁶ Thus at present there is very little research on which to base plans for the preservation and continued usability of digital documents; and most of the existing work in this subject area tends to be conceptual or theoretical.

Consequently there is an immediate need for research that considers real world operating constraints. Empirical efforts to apply and test digital preservation approaches can provide valuable new theoretical insights and constructs while demonstrating trial systems with capabilities that can subsequently be implemented in government agency pilot projects.

⁷⁴ An account of early efforts to address the implications of uses of digital media in official activity for archives and records management in The Netherlands is available in T. K. Bikson and E. J. Frinking, *PRESERVING THE PRESENT*, The Hague: Sdu Publishers, 1993.

⁷⁵ Agencies in many countries are now attempting to clarify what is legally or institutionally required to establish the official status of digital interactions (see, for example, M. Hedstrom and F. Blouin, Jr., *ELECTRONIC RECORDS RESEARCH AND DEVELOPMENT*, Ann Arbor MI, 1997 or www.si.umich.edu/e-recs/; D. Roberts, "Defining Electronic Records, Documents and Data," *Archives and Manuscripts*, Fol. 22(1), 1994, pp. 14-25.

⁷⁶ Two recent conferences, however, have focused attention on preservation problems associated with digital records: *Digitale Duurzaamheid*, supported by the Ministry of the Interior and the National Archives of The Netherlands, Rotterdam: 9 April 1998; and *The XXXIIIrd International Conference of the Round Table on Archives (CITRA)*, Stockholm: 9-12 September, 1998.

In a 1998 paper, John McDonald contends:

In light of the growing number and significance of record keeping issues being faced by modern organizations, archives ... may be confronted by anxious organizations demanding that they come up with effective and relevant record keeping solutions and that they do so without delay.⁷⁷

This view is echoed in a current review of archival practice, which notes that “increasingly, organizations and individuals are demanding guidelines and standards” to achieve authentic and reliable digital documents whose evidential and information value is guaranteed over time.⁷⁸

The concept of good governance in open societies has long been based on the availability and accessibility of trustworthy records. Traditionally the role of the archives has been to provide the knowledge and techniques required to assure those ends. The proposed testbed would help the archives fulfill that role in the digital government era.

C.2 Research Questions

The overall goal for the development of a testbed is to generate the knowledge and techniques needed to carry authentic, understandable and usable digital records through time. Key research questions for the testbed to address include the following.

- * What lessons can be learned from emerging digital records preservation practices in other institutions, including agencies of other governments as well as large corporations and computing centers, that may help inform testbed design and research activity and—ultimately—the development of operational digital preservation functions for the National Archives of The Netherlands?
- * What record-critical document attributes (e.g., appearance attributes, behavioral attributes) are unique to or markedly different in digital media?
 - How, if at all, do these attributes differ as a function of document type?

⁷⁷ J. McDonald, “Current Records, Access and the International Archival Community,” *CITRA Proceedings*, session 3, Stockholm: 9-12 September, 1998.

⁷⁸ J-P. Wallot, “A Look at Archival Practices,” *CITRA Proceedings*, session 2, Stockholm: 9-12 September, 1998; requests for digital preservation guidelines in general, and for help in bridging the gap between preservation models and successful preservation practices in particular, are documented in an empirical survey by M. Hedstrom and S. Montgomery, *DIGITAL PRESERVATION NEEDS AND REQUIREMENTS IN RLG MEMBER INSTITUTIONS*, Mountain View CA: Research Libraries Group, 1998 or www.rlg.org.

- How, if at all, do these attributes differ as a function of the specific software in which the document is encoded?
- * What kinds of technical or other metadata are necessary and sufficient to describe these record-critical characteristics of digital documents?
 - How, if at all, do these metadata requirements differ as a function of document type or critical attribute type?
 - How, if at all, do these metadata requirements differ as a function of the specific software in which the document is encoded?
 - At what levels (e.g., the record item, the series, the record-keeping system) and in what ways (e.g., encapsulation, linking) can metadata be reliably associated with digital documents? What are the comparative merits and drawbacks of different choices?
- * How can the validity of alternative digital preservation approaches be corroborated?
 - What validation criteria bear on both evidential and informational authenticity of digital documents? What validation criteria are unique to determining evidential vs. informational value of digital documents?
 - How, if at all, do the answers differ as a function of document type or critical attribute type?
- * Which digital preservation approaches, if any, will meet empirical tests of feasibility and validity? What are their comparative strengths and weaknesses as techniques for preserving authentic digital records?
 - How, if at all, does the effectiveness of alternative approaches differ as a function of document type or critical attribute type?
 - How well do selected sets of technical metadata serve to support alternative preservation approaches?
 - What is the relative technical ease or difficulty of applying alternative preservation approaches to digital documents?
 - How do alternative technical approaches differ, if at all, with respect to ease of viewing, understanding and (re)using preserved digital documents?
- * What techniques can be developed to support the automated acquisition of digital documents, along with their critical attributes and extant metadata, from existing digital recordkeeping systems for ingestion by archival preservation systems?

- * What are the implications of alternative digital preservation approaches for the design of archival digital repositories?
- * What are the implications of alternative digital preservation approaches for relationships to extant archival management systems?
- * What are the implications of alternative digital preservation approaches for coordination between digital and nondigital archival records?
- * How are alternative digital preservation approaches expected to differ in cost effectiveness?
 - Can the expected costs of digital preservation be factored (e.g., into initial costs associated with ingestion and preparation vs. long term maintenance and (re)use)?
 - Do alternative approaches to digital preservation differ in how costs are distributed among such factors over the digital preservation time horizon?

As may be evident, the questions suggested above are ordered in that research answers to those earlier in the set are expected to prove helpful, and sometimes necessary, for research that addresses the later items. However, the set of questions should not be regarded as complete. Testbed activity is likely to generate new research questions and improved formulations of initial questions over time.

C.3 Scope

The testbed should provide an environment for conducting experiments designed to answer predominantly technical questions like those outlined above about the preservation of digital records appraised as having archival value. At least initially it is not intended to represent in microcosm the actual workflow associated with an operational digital archives. On the other hand, the experiments it carries out should be designed as veridically as possible, reflecting the real world conditions under which digital records are generated by government agencies so that they will be able readily to incorporate the testbed's findings into their own plans and pilot projects.

The testbed, then, will begin its activities with a fairly small sample of digital records; and the sample will be chiefly textual, omitting multimedia, interactive and executable information objects from its initial research agenda. However, the starting sample will be drawn from the record-creating work of a few cooperating agencies and will include a limited number of types of digital documents in frequent contemporary use (e.g., e-mail with attachments or reports with tables and figures).

Similarly, although the testbed will aim at building a metadata structure that is sufficient to encompass a broad range of digital information objects, it will begin by

attending to a minimum set of basic technical attributes assumed to be necessary for the preservation of the key characteristics of the initial document sample. Finally, although the testbed should be suitable for exploring varied preservation strategies as they emerge, in the beginning it will focus on just two proposed options: migration and emulation.

These kinds of boundaries form the basis for the recommended start-up infrastructure outlined in Section C.6 below. However, the iterative spiral approach to development and experimentation suggested here requires the testbed infrastructure to be extensible as the scope and complexity of its activities increase. As it matures, therefore, the testbed might evolve from one that targets specific technical problems to one that approximates the archival preservation workflow as an entire process.

C.4 Tasks

In what follows, the development of the testbed is described in a series of seven suggested tasks and one preparatory task. It is assumed that the tasks would be performed on a repeating basis, with later iterations to be informed by results from previous efforts. Iterative spiral development processes of the sort recommended have been well documented in previous literature and are especially suitable for testbed tasks such as those outlined here.⁷⁹ Further, although they are in many respects interdependent, the tasks are each expected to yield some outcomes of independent interest.

In the task discussions below, a brief summary is given first. Then the task is more fully explained and relevant prior research, if available, is cited.⁸⁰ Last, expected task outputs are described.

⁷⁹ D. Mankin, S. Cohen and T. K. Bikson, *TEAMS AND TECHNOLOGY: FULFILLING THE PROMISE OF THE NEW ORGANIZATION*, Boston MA: Harvard Business School Press, 1996; T. K. Bikson, S. Cammarata, S. A. Law and T. West *UN/UNESIS Phase I Report: Plans for the Progressive Implementation of UNESIS*, Santa Monica: RAND, DRU-910/4-UN, 1995; P. Seybold, "How to Leapfrog Your Organization into the Twenty-First Century" in *HIGHLIGHTS FROM PATRICIA SEYBOLD'S TECHNOLOGY FORUM*, New York: Patricia Seybold Group, 1994, p. 2; B. Boehm, "A Spiral Model of Software Development and Enhancement," *Computer*, Vol. 21, No. 5, May 1988, pp. 61-72.

⁸⁰ We were able to locate very little previous research that would be directly helpful for the design and implementation of a digital preservation testbed. (As explained earlier, most of the prior literature in this field has been exclusively conceptual rather than empirical or applied.) The reports most useful for our purposes, in order of descending importance, are the following: I. Starostine and C. Sevy, *Non-Proprietary File Format — ERM Pilot Project Final Report*, Washington DC: International Monetary Fund (IMF), Secretary's Department, 1997; F. Fonseca, P. Polles and M. Almeida, *Analysis of Electronic Files of Bank Reports*, Washington DC: The World Bank, Information and Technology Services, 1996; A. R. Heminger and S. B. Robertson, *Digital Rosetta Stone: A Conceptual Model for Maintaining Long-Term Access to Digital Documents*, 1998; and I. Fønnes, "Review of Archival Practice," *CITRA Proceedings*, session 2, Stockholm: 9-12 September, 1998.

Preparatory Task Survey the state of the art in digital preservation

As preparation both for the development of an operational testbed and for the specific choices to be made in carrying out an initial set of digital preservation research tasks, it is appropriate to survey lessons learned by other organizations. Considerable experience with varied aspects of digital preservation (e.g., guidelines for storage media, information security, migration processes, metadata requirements) is likely to reside in large commercial companies or computing centers as well as in agencies of other governments. Identifying such organizations and learning from best practices will enable the testbed to start with state-of-the-art knowledge and, over time, to advance it.

Such a survey should be conducted extensively when testbed development begins and the findings should be updated at least annually. The technologies involved in digital documents are undergoing rapid and continuing change, as are associated work practices in organizations. Both are likely to have significant, if not yet foreseeable, effects on preservation techniques and tools. Although the primary purpose of this preparatory task is to inform testbed design and research activity, the results would be expected to be of considerable interest across the international archives and records management communities.

Task 1 Select an initial sample of test documents and devise an acquisition method

The first task of the testbed is to create a limited but varied sample of digital documents. These documents will constitute the base material on which subsequent tasks are to be carried out.

Because this is envisioned as an iterative task, the initial set of documents chosen for study need not be large. It is more important that the sample be varied, including representatives of different types of documents in common use among agencies (e.g., memoranda, correspondence, minutes, reports). The typology offered in a recent UN/Information Systems Coordinating Committee report provides a helpful foundation for making the selection.⁸¹ While the starting sample should include some compound documents, we recommend that the components be limited initially to tables, figures, images, attachments and the like, excluding multimedia and interactive objects from the initial effort.

⁸¹ T. K. Bikson, "Strategies for Implementing Document Management Technology," *Annex II: Report of the Task Force on Document Management Methodology, Information Systems Coordinating Committee (ISCC)*, New York: United Nations, Administrative Committee on Co-ordination (also available from RAND Corporation, Santa Monica CA, Reprint Series, publication RP-704, 1997. The typology is subsumes document types under critical business processes common to governmental and nongovernmental organizations. A useful discussion of issues in moving from organization-specific document typologies to typologies that are "global" (or, that can be used effectively across organizations) is found in McKemmish and Parer, 1998, op cit, note 3.

In the starting sample, moreover, it is important to include multiple instances of each chosen type from the “real” digital documents of two or more agencies. That is because, even when two organizations make use of a common hardware and software platform and rely on similar types of documents, different work practices are likely to result in different issues for the retention of usable digital records that would not have been anticipated from a theoretical standpoint. Such issues are likely also to emerge within organizations as well, if they comprise quite different primary business processes or exhibit considerable unit-level autonomy.

Having decided on the nature of the document sample, the next step in task 1 is to devise a method for acquiring the desired items. The most straightforward procedure would be to capture documents of the desired types from active digital records management systems in cooperating agencies. Unfortunately, even though most documents are prepared and shared in digital form, many agencies still continue to treat only paper-based documents as official record material.

As a result, it may not be possible to acquire digital documents directly from a digital system in which they have been registered as records. Other options for acquiring a veridical sample include the following. First, digital versions of documents may be acquired at the final stage of preparation (i.e., from the originating departments), before they are transferred to a printing or publications department. Obtaining digital documents in this way will require copying them to testbed storage devices, potentially risking the loss of some system environment or business context information while introducing an added transaction (copying). An internal study carried out by the International Monetary Fund (IMF) nonetheless found this kind of acquisition procedure to be relatively robust.⁸²

In contrast, an experimental effort to select a sample of digital documents for an internal World Bank study found such a strategy not to be viable. From a set of over 135 recent project reports for which authoring units believed they had final digital versions, complete digital material could be found only for about 50 percent. In that subset, however, just 22 percent had been created in their entirety in a single digital file; the remainder comprised multiple individual files, averaging about 7 files per document. Further, assembling the multi-file documents was not possible without using the published report as a guide.⁸³ Consequently, the Bank at present captures its digital documents by scanning the printed versions.

Such a procedure is not recommended for purposes of acquiring a testbed document sample because it would eliminate both native formatting and contextual information while introducing more additional transactions (printing and scanning). Thus if it is not feasible to obtain copies of stored final digital document versions (following the

⁸² Op cit, note 80. This IMF project as well as The World Bank project also cited in note 80 illuminate in significant ways the kinds of unanticipated lessons that can be learned in the attempt to apply theoretically well developed constructs and processes.

⁸³ Op cit, note 80.

IMF procedure described above), we recommend prospectively selecting such a sample and requesting that copies be transferred to the testbed at the time they are being sent to printing or publications departments, or to other units or organizations (e.g., via transfer to a shared file server or by electronic mail). The viability of the latter approach presupposes the willingness of cooperating agencies and selected authoring units to take part in procedures that might make added demands on their time or other resources.

While defining and acquiring an initial sample of digital documents serves as the basis for a number of subsequent testbed activities, it will also provide the archives with limited baseline information about the kinds of issues it is likely to face as agencies begin to submit digital material for long-term preservation. Iterative additions to the testbed's digital document sample will thus help inform the archives' efforts to articulate criteria for acceptability of digital record submissions ("ingestion" criteria).⁸⁴

Task 2 Analyze key attributes of digital documents to determine those that are critical to their preservation and use as digital records with archival value

For purposes of carrying out the second task, we recommend restricting the analysis to the sample of documents selected in the first task while bearing in mind that the sample does not exhaust the range of attributes of potential importance to digital records and archives. This approach is intended to keep the task empirically grounded but extensible.

Further, we suggest focusing the task on characteristics of documents that are either unique to or significantly different when manifest in digital media. Embedded cell formulas in spreadsheets and active links within or between documents, for example, are attributes without counterparts in the nondigital world. On the other hand, while contextual information (e.g., time of creation, routing information) characterizes both paper and digital records, it may be captured or represented quite differently in digital systems.

As a starting point, we suggest considering five categories of document attributes: content, structure, context, appearance and behavior. Among these, the first three probably have the greatest number of counterparts in paper media. Even in these categories, however, digital documents raise some new issues. A considerable amount of content, structure and context information is borne by the physical form or location of paper documents. For instance, the structural relationship of "attachment" might

⁸⁴ Ingestion processes, and specifically the necessity to identify and take into account specific software dependencies of files in all cases where specific software may be required either to retrieve or display any record in the acquisition of records, are discussed briefly in the work of Reagan Moore: see A. Rajasekar, R. Marciano, and R. Moore, *Collection-Based Persistent Archives*, San Diego Supercomputer Center, <http://www.sdsc.edu/NARA/Publications/OTHER/Persistent/Persistent.html>.

be represented by physical connection; the origin, the order of records in a series, and other contextual information may be shown by the physical location of a file cabinet and the placement of items within it. More generally, what was represented in physical ways and reinforced by routine work practices in the world of paper records is likely to be manifest quite differently in digital systems. These differences need to be analyzed and their implications for preserving the informational and evidential value of digital documents must be assessed.

We have put attributes related to document appearance and document behavior in separate categories for purposes of focusing task 2, even though it might be possible to construe them as falling within one or another of the previous three groups.⁸⁵ Characteristics of document appearance raise special problems because digital records are not readable without appropriate hardware and software; consequently, decisions about the appearance attributes that are significant from an informational and evidential perspective have direct implications for the digital retention and re-use strategies to be explored in the testbed. At minimum, there must be testbed software that enables the viewing of digital documents created elsewhere; beyond that, task 2 also involves deciding which attributes of appearance are critical to the record value of documents of different types. Is bold or italic appearance important, for instance, or is it only important to be able to view a particular text string and know that it was highlighted? Similar questions could be addressed to color coding and many other appearance-attributes of documents.⁸⁶ (Answers to these questions will affect the kinds of validation tests to be generated in task 4).

Behavioral attributes of digital documents also pose unprecedented issues. Here the term “behavior” is used to reflect the interactive properties of digital records—the characteristics that allow them to be not just “viewed” but also to be “used” or “processed” in their native software environments.⁸⁷ For example, documents may be subject to full-text search or their active links pursued; “what if” queries can be conducted with data in spreadsheets, while relational databases will return answers to indefinitely many well-formed queries. Among the behavioral attributes of different types of digital documents, then, completion of task 2 requires determining which of them must be preserved in order to meet requirements for retaining informational and evidential value. This component of task 2 will be difficult for the initially selected document sample because it has no clear precedents in traditional archival methodology. It is expected to become more even problematic as subsequent

⁸⁵ Most sources treat document properties in terms of the first three categories (content, structure, context). For testbed purposes we believe it is appropriate provisionally to give special attention to attributes of appearance and behavior by creating separate categories for them. It is possible that when a better developed and validated conceptual framework for digital records is available it will be possible to produce a more parsimonious categorization of record-relevant document attributes.

⁸⁶ An instructive example of the critical role that color attributes played formerly in the paper document context and currently in the digital document context for official government agency work in Germany is found in W. Prinz and S. Kolvenback, “Support for Workflow in a Ministerial Environment,” *Proceedings of the Conference on Computer Supported Cooperative Work*, New York: Association for Computing Machinery, 1996.

iterations of task 1 incorporate documents with multimedia components or components that are themselves interactive.⁸⁸

Findings from task 2 will provide a foundation for the development of validation tests discussed in task 4. Additionally they will provide a preliminary indication of the range of unique attributes that must be preserved in digital documents as well as the extent to which these attributes are (a) appropriately assigned to the item level, the series level or both, (b) consistent within similar document types selected from different agencies, and (c) common across different document types, whether selected from the same or different agencies.

Task 3 Represent needed metadata and determine appropriate methods for associating them with digital documents

As with task 2, we also recommend initially limiting task 3 to metadata for the chosen sample of testbed documents and focusing new research on aspects of metadata definition and attachment that are likely to be new or different in relation to digital records. In particular, extant metadata associated with the selected documents in the organizations' recordkeeping systems should be acquired for the testbed. However, these metadata will most likely have to be extended to reflect the technical properties of digital records and their software and hardware environments. Moreover, task 3 should consider technical aspects of capturing and attaching these metadata.

Three prior projects provide a useful starting point for carrying out this task: the previously described IMF study; the Australian framework for standardizing recordkeeping metadata related to "document-like information objects"; and the work of the United Nations Information Systems Coordinating Committee (ISCC) task

⁸⁷ Behavioral attributes of digital records have been given greatest attention when relational databases are under consideration (see for instance, Fannes, 1998, op cit, note 80, and Bikson and Frinking, 1993, op cit, note 74, on archival approaches in Sweden). However, preservation of behavioral attributes in statements of criteria for preservation of digital records are not typically restrictive with respect to type of record. The US Assistant Secretary of Defense for Command, Control, Communications and Intelligence, for instance, has included among "mandatory requirements" for electronic records management the following (November 1997):

C2.2.4.2 Ability to Read and Process Records. Since [Records Management Applications] are prohibited ... from altering the format of stored records, the organization shall ensure that it has the ability to view, copy, print and, if appropriate, process any record stored."

US Department of Defense *Design Criteria Standard for Electronic Records Management Software Applications* (DoD 5015.2-STD) can be found at <http://jitic.fhu.disa.mil/rec/mgmt/>.

⁸⁸ A recent UN/ISCC report provides evidence of the growing use of multimedia, interactive, and dynamic applications-like documents among United Nations Agencies (T. K. Bikson, *Roadmap to Electronic Document Management in United Nations Organizations*, New York: United Nations Information Systems Coordination Committee (ISCC), Administrative Committee on Coordination, ACC/1997/ISCC/5, 1997). Among the behavioral properties of documents described in that report are interactive maps and project reports generated on demand with dynamically incorporated current budget material as well as instructional or training documents with embedded quizzes and tutorially-designed multimedia clues to correct answers.

force on document management technology.⁸⁹ In each case, the projects reviewed and adapted sets of metadata from the Dublin Core elements, the Pittsburgh metadata elements required for “business-acceptable communications,” or both. The ISCC task force, in particular, recommends a minimum critical set of metadata specifications for mandatory use by all international organizations. It includes elements from both the Pittsburgh and Dublin sets, emphasizing the former for its attention to the evidential value digital documents. The IMF study builds on those recommendations, carefully elaborating metadata elements related to technical properties. The ISCC metadata set, like its two predecessors, is intended to be generic and extensible.

Task 3, then, might begin by adapting this set of metadata elements and elaborating the technical specifications to represent the properties of the digital documents in the testbed sample. With an appropriate metadata set established, the next steps in task 3 involve considering how such metadata should be associated with the documents being preserved and whether they might be captured directly along with the documents from their native system environments. Metadata might be associated with digital documents, for instance, at the system level or at the level of the individual record item; in the latter case, they may be linked to the document or encapsulated with it or both. Although encapsulation seems the safest way to ensure that needed metadata remain with the records, some descriptive material must also be retained in contemporary systems to enable search and retrieval of digital documents in archival repositories. The testbed will enable archival researchers to explore advantages, disadvantages and tradeoffs among differing approaches to metadata attachment.

Ideally, the testbed will also enable experimentation with procedures for the automatic generation or capture of at least some needed technical properties from documents’ native system environments at the time of acquisition. However, the capacity to carry out these explorations is dependent in part on the acquisition options revealed during task 1 and also on the willingness of agencies to cooperate in research activities that could potentially perturb routine operations.⁹⁰ Consequently it may be best to defer experimentation with direct capture of technical metadata until later rounds of testbed activity.

In any event, a valuable product from the testbed as task 3 is completed will be a preliminary set of core technical metadata for official digital documents, modified to reflect the institutional context and government practices of agencies of The Netherlands. And, on completion of task 3, the testbed will have become a functioning

⁸⁹ “Proposed Mandatory Set of Core Metadata,” *Annex IV: Report of the Task Force on Document Management Methodology*, New York: United Nations Information Systems Coordination Committee (ISCC), Administrative Committee on Co-ordination, ACC/1997/ISSC/4, 1997; Starostine and Sevy, 1997, op cit, note 80; McKemmish and Parer, 1998, op cit, note 3.

⁹⁰ Starostine and Sevy, 1997, op cit, note 80 (the study warns about the pitfalls, both practical and methodological, of interfering in course of a primary business process while trying to capture veridical data from and about it).

research environment; that is, it will support experimentation with alternative methods for attaching metadata to a representative subset of digital documents and examining the results.

Task 4 Devise a set of validation tests, ideally based on results of task 2 and appropriate independently of any particular preservation method

Validation tests here are conceived as tests of whether or not a given procedure can be said to have preserved the original digital record (see the earlier discussion of what is meant here by “digital original”). Put another way, validation tests serve to determine empirically whether all properties that bear on evidential and informational value of digital records remain intact and whether, or to what extent, their authenticity has been preserved.

The job of task 4, then, is to construct a suite of tests for determining whether the critical attributes of digital documents identified in task 2 exist unaltered in a system environment that differs from the ones in which they were created. The focus, again, should be on attributes that are likely to present new or different problems in relation to digital records. As we have said, we expect these to be attributes of appearance and behavior.

An instructive example comes from a 1993 interview carried out with one department head in the Ministry of Agriculture, Nature Management and Fishery in The Netherlands.⁹¹ The department had recently converted a sizable data set from one SQL-compliant relational database to another. It took only about 6 weeks to transfer the numeric data elements and confirm that the contents had not been altered in the process. However it took another 6 months to test the embedded database formulas and relationships. That part of the testing process surfaced a great many instances in which relationships were severed or altered; and formulas did not consistently behave in the new environment as they had in the old.

A highly valuable function of the envisioned testbed, then, will be to devise validation tests at a time when the results of the tests done on migrated, emulated or otherwise preserved digital documents can be compared with those same tests run on the documents in their native software and hardware environments. It is impossible to overestimate the value of determining, objectively and in advance, whether a proposed preservation strategy will succeed.⁹²

Efforts are well underway in Scandinavian countries to produce exhaustive logical descriptions of the behavior of relational databases.⁹³ These could provide a productive starting point for the development of a suite of validation tests for records of that type. Our search of relevant literature did not reveal comparable progress

⁹¹ Bikson and Frinking, 1993, op cit, note 74.

toward the production of logically complete descriptions of the behavior of other document types (e.g., text with attachments, active links or embedded objects).

Appearance attributes, in contrast, have to do with what now is often referred to as “look-and-feel,” although the results of task 2 should yield more precise accounts of the appearance characteristics of interest for informational and evidential value. However, it is not clear how to impose objective criteria for record-faithful rendering even after the critical appearance attributes have been defined. Initially the testbed might rely on an advisory panel comprising both domain experts and knowledgeable professionals in the field of digital records and archives. Systematically collected expert judgments about the fidelity of appearance attributes required for the retention of informational and evidential value could provide the starting points for iterative efforts to define rendering validation tests.

Activities undertaken to complete task 4 will generate means for determining whether currently proposed preservation methods can succeed in providing for the long term retention and use of digital records. In the process, they will help inform ongoing debates about viewing standards and related software for access and use. These issues become increasingly important as multimedia and interactive objects are more frequently being incorporated in Internet and Web-based documents.⁹⁴ Further, completion of task 4 should contribute to theoretical understanding of the nature and role of appearance attributes in conceptions of digital records.

Task 5 Carry out the sequence of activities required for proposed preservation methods; document in detail the processes involved in

⁹² Determining that any given preservation method works, in advance, while the older “native” technology that produced it is still available is critical for several reasons. First, after a particular software format has become obsolete, it will be much more difficult to make a sound assessment of the extent to which the chosen procedure in fact preserves critical informational and evidential attributes of records created in the older software. Second, only if such preservation issues are addressed in advance is it likely that they will be handled systematically. The report by Hedstrom and Montgomery (1998, op cit, note 78) reveals that in the US, at present, preservation procedures such as migration or conversion are typically undertaken on an ad hoc basis (e.g., as a function of technical departments’ decisions to undertake system upgrades) quite independently of any considerations based on archival methodology. Third, systematic advance attention is probably the best way to cope with the fact that the preservation problem is a “hidden” one—the unusability of old digital documents is unlikely to be discovered unless and until the records are needed, which may be too late for all but the most heroic and costly of measures (T. K. Bikson, “Digital Documents: Technology Trends and Organizational Responses,” *Programma Digitale Duurzaamheid*, Rotterdam: 9 April 1998; see also Hedstrom and Blouin, 1996, op cit, note 75).

⁹³ See the discussion in note 87. Presumably the sort of work done to generate logically complete and generic (nonproprietary) descriptions of the behavior of relational databases might also be attempted for spreadsheets with comparable success. However, as the main text of this report argues, it is not clear that such methods would generalize to other types of records (e.g., hypermedia text documents, object-oriented databases, and, by extension, to paradigms yet to emerge).

⁹⁴ Bikson, 1997, op cit, note 88; Bikson, 1998, op cit, note 92.

their implementation, noting the choice points, options considered and decisions made along the way

Tasks 1 through 4, while yielding archival research information of independent interest, also prepare the foundation for conducting tests of alternative preservation methods and comparing the outcomes. The results, based on a multi-agency sample of diverse digital documents and a standardized suite of validation tests, would provide hitherto unavailable information about digital document preservation that should contribute basic building blocks for twenty-first century archival methods.

As explained elsewhere in this report, the most widely accepted approach to the problem of digital durability (or, to the problem of obsolescence of native software and hardware) is migration or conversion. These terms embrace not only the re-copying of digital records to new media (sometimes called “refreshing”) but also their continuous conversion to new generations of software and hardware as the native environments of their creation and use obsolesce but the standards that define them remain constant.⁹⁵

Elsewhere we have argued that standards-based migration or conversion is problematic for at least two kinds of reasons.⁹⁶ One has to do with standards. While ISO standards have been adopted for many widely used digital technologies, they are typically implemented in proprietary software along with other vendor-specific features. Because appearance and behavior of digital documents are much more likely to be affected by such vendor choices than are content, structure or context attributes, it is unclear whether migration or conversion will preserve them. Further, the relatively rapid introduction of new digital information and communication paradigms on the one hand (e.g., object-oriented systems) and the relatively slow development of standards and thus of standards-compliant software products on the other (e.g., Adobe’s PDF is not yet a standard), suggests that relying on standards for preservation is—at least in the near term—an “act of faith.”⁹⁷ Because there is no end in sight to advances in digital information and communication technologies, it is unrealistic to expect a defined set of standard forms to meet future archival

⁹⁵ A. Smith, “Digital Archiving Models,” *CITRA Proceedings*, session 5, Stockholm: 9-12 September, 1998; see also Fonnes, 1998, op cit, note 80, where it is pointed out that migration is the “second most widely accepted strategy after doing nothing.”

⁹⁶ Rothenberg, 1995a, op cit, note 2; J. Rothenberg, “Survival of the Digits,” *Communicating Business*, Issue 4 (spring), London, Forward Publishing, 1995b; Rothenberg, 1999, op cit, note 31.

⁹⁷ Bikson, 1997, op cit, note 88. This point was underscored in an interview with Ken Thibodeau, head of the US National Archives and Records Administration’s Modern Records Division, for the present project. Responding to a question about the likely durability of PDF, given that it is a de facto standard and that its proprietary software is distributed without cost, he replied “How do you know that Adobe is not the Ashton-Tate of the 1990s?” His point is that, for any proprietary software, it is important to take into account—among other concerns—the longevity of the firm that produces it. Although Ashton-Tate produced the most widely used relational database software in the 1980s, it is no longer in business. Any preservation solutions that depend on the continued existence of proprietary firms (e.g., to provide upwardly compatible software, to guarantee migration or conversion paths and/or to produce open standards-compliant products) at this point probably depend more on hopes than on methods.

preservation needs. Finally, while migration may appear relatively straightforward, it requires unique, specialized treatment of each different document type, as well as individualized application to every document whenever obsolescence necessitates migration.

Emulation offers a contrasting approach with a number of theoretical advantages.⁹⁸ Emulating obsolete computers on future computers would allow running the original software that created or viewed a digital record, even after that software becomes obsolete in the future. Furthermore, emulation would require no special treatment of different document types and no processing of individual documents, other than saving their bit streams, though it would require saving software and associated documentation as well as descriptions of popular types of computers sufficient to allow generating future emulators for them. Moreover, while paradigmatic or even simply evolutionary changes in software applications and tools may be difficult to foresee, hardware changes are easier to anticipate. They have evidenced steadily improving price-to-performance ratios for several decades.⁹⁹ It now goes without saying that most of the computers on individual users' desktops can outperform the mainframe computers that formerly served all of an organization's users. More generally, it is reasonable to infer that newer hardware will always be capable of emulating the performance (as well as the functionality) of older hardware. Further, there are many successful examples of emulation, including (for instance) MAME, the Multiple Arcade Machine Emulator, which allows users to access and use now-obsolete game software. On the other hand, there are no generalized and formal specifications for emulation; and most organizations have little or no internal experience with emulation.

Against this background, a first priority for the testbed should be to design and conduct trials of migration/conversion and emulation strategies, using the initially selected sample of digital documents as test material. These efforts can be construed as separate series of subtasks, as outlined below. They can be carried out concurrently or sequentially, depending on the resources available to the testbed.

An emulation trial, for instance, can be construed as involving the following set of subtasks.

- Develop hardware emulator specifications, based on the range of behaviors that are inherent in the native forms of the document types and instances in the testbed sample.

⁹⁸ Rothenberg, 1995a and 1999, op cit, note 96.

⁹⁹ US National Research Council, *FOSTERING RESEARCH ON THE ECONOMIC AND SOCIAL IMPACTS OF INFORMATION TECHNOLOGY*, Washington DC: National Academy Press, 1998.

- Determine viewing software requirements and select a viewer capable of rendering the appearance attributes of document types and instances in the sample.
- Include hardware emulator specifications and viewer requirements along with the original software in the technical metadata associated with the document (in addition to any other metadata generated by task 3); encapsulate the software with the document and other metadata but also retain emulator and viewer information in a form readable by contemporary systems.
- Produce a running emulator on contemporary hardware that is different from the documents' original hardware environments. In early testbed trials, this process should be expected to be done largely "by hand," although it may be aided by the creation of intermediate forms, translators or other tools.
- Port the chosen viewing software to the emulated hardware environment, if necessary. Ideally, this step will not be necessary. However, because the emulator specification and the emulator itself will be prototypes, some adaptations may be required to make the viewing software work during early testbed trials.
- Bring up the native software in the emulated hardware environment and regenerate the original digital documents.

For purposes of trying conversion approaches to preservation, we have not located a clearly delineated yet generalized series of steps. The IMF study, although restricted to a smaller sample of document types drawn from one agency, nonetheless presents a model that could be useful for organizing trials based on SGML conversion. Subtasks might include the following.¹⁰⁰

- Prepare DTDs for the digital documents in the testbed sample, first generating those that are most generic (e.g., for standardized table formats), then those that are specific to document types (e.g., titles for memoranda vs. working papers), and finally those that are unique to particular documents.
- Identify appropriate software for converting existing digital documents to SGML files (e.g., Omnimark).

¹⁰⁰ This hypothetical series of summarized steps is based on the account of the IMF project to create non-proprietary digital files for long-term retention of and access to records (see Starostine and Sevy, 1997, op cit, note 80). It is intended as an example only; it does not do justice to the detailed explanation presented in their report.

- Determine the processes and mechanisms required for successful conversion (e.g., in the IMF study it proved to be necessary first to parse through the document-unique DTDs and then to parse the shared elements; additionally it proved to be necessary to convert original WordPerfect documents to RTF before the conversion to SGML files would work).
- Carry out the requisite conversion routines to produce SGML files from original documents (for preservation).
- Encapsulate technical and other metadata with the resulting SGML files; retain in contemporary or human-readable media the information needed to render the SGML files viewable.
- Create HTML document versions for viewing, searching and other uses of the preserved documents.

Emulation and SGML-based conversion are used here as initial examples of preservation approaches that could be prototyped and assessed in the testbed. They are suggested as starting points because of their popularity or their promise. PDF might also be subjected to a conversion trial like the one outlined above. Use of PDF is becoming more common in organizations as an alternative to SGML for enabling the retention and reuse of digital documents because of its superiority for representing their appearance attributes. The International Civil Aeronautics Organization (ICAO), for instance, has adopted PDF for its digital document management system.¹⁰¹ On the other hand, because PDF remains a proprietary product of Adobe (although readers for it are distributed without cost) and because it is not an official standard (although it is expected to become one), it does not presently meet U.S. requirements for archival media.¹⁰²

Besides standards-based conversion and emulation, other approaches could be explored following comparable procedures. For example, the “Digital Rosetta Stone” model outlined by Heminger and Robertson could become a candidate for the testbed.¹⁰³ The weakness of the model at present is the lack of any methods or precedents for specifying all the “meta knowledge” necessary to reconstruct and interpret a digital document once the underlying preserved bit stream had been regenerated. A substantial amount of effort would be required to operationalize the model. On the other hand, the testbed could also be used to test the capability of generic relational database models to preserve all the informational and evidential characteristics of database records produced using recent database software.¹⁰⁴

¹⁰¹ Bikson, 1997, op cit, note 81

¹⁰² See also note 97

¹⁰³ 1998, op cit, note 80.

¹⁰⁴ See note 80.

These approaches have been relatively well specified; however, we have not used them as examples because of uncertainty about their generalizability to other types of digital records. In any case, the testbed research procedures—and validation tests in particular—are designed to be as generic as possible and neutral with respect to alternative preservation approaches. As proposed new approaches (e.g., universal viewers) become sufficiently well developed, it should be easy to incorporate them into the testbed research agenda.

Task 5 is, to be sure, ambitious. However, even the first iterations would produce significant empirical data—now lacking—about the feasibility of the most promising approaches thus far proposed for preserving digital records.

Task 6 Perform the set of validation tests developed in task 4 on the “preserved” digital documents that result from task 5

While completion of task 5 will yield detailed findings about the feasibility of proposed digital preservation procedures, task 6 will test the outcomes against previously established success criteria. To our knowledge, no formal tests of this type have been carried out. On the other hand, our literature review surfaced a number of less formal approaches.

For instance, dedicated arcade users find that MAME provides the look-and-feel of older computer games while faithfully reproducing their behavior. More serious and successful single-case uses of emulation are also described in a 1999 Rothenberg report.¹⁰⁵ On the other hand, IMF’s project to test the conversion approach yielded mixed results. The research team concluded that SGML conversion satisfied needs to preserve content, structure and context characteristics but did not retain attributes of appearance and behavior. Thus the procedure did not seem to support future use of preserved documents.

Carrying out task 6 would enable these kinds of conclusions to be drawn on a previously validated and systematic basis. At present there are no objective yardsticks by which to assess the effectiveness of any of the proposed methods for retaining usable digital records. Further, findings from the validation tests could be compared across different preservation approaches. For document types where alternative preservation options provide equally successful outcomes across critical attribute categories, decisions about preservation approaches might turn on cost or ease of implementation.

Task 7 Store the “preserved” digital documents and metadata plus the information required to locate and re-use the documents; search for

¹⁰⁵ Rothenberg, 1999, op cit, note 31.

and retrieve a subset of desired documents; repeating tasks 5 and/or 6 on the subset.

Producing the output from task 7, while useful for corroborating prior testbed procedures, will also provide information about the characteristics that make it difficult to locate and use saved digital documents. While these task results will be of major importance for archives, they are less important for testbed purposes. The chief contribution of task 7 is to call attention to the role of the repository and the need to assure the viability of search and retrieval mechanisms for digital documents of enduring informational and evidential value.

C.5 Results

As the foregoing discussion has indicated, we believe that each of the proposed testbed tasks is capable of making independently valuable contributions to theory and practice in the field of digital archives and records. Collectively they comprise a research agenda addressing vital needs for preserving the digital documents that will become the major media for government agency interactions in the twenty-first century.

The results of the suggested testbed agenda should provide iteratively improved answers to questions about the record-critical document attributes (especially, appearance and behavioral characteristics) that are unique to or markedly different in digital media. It will augment current understanding of metadata requirements, particularly requirements related to technical or other special metadata that are necessary or sufficient to describe digital records. Further, it will provide a general method for corroborating the validity of alternative digital preservation approaches. Finally, it will determine which digital preservation approaches, among those now available or soon to emerge, will meet empirical tests of feasibility and validity. And, as the testbed yields results deemed ready for application, its findings should yield data useful for addressing a range of questions bearing on the development of fully operational preservation systems and procedures.

The conclusions should be of profound interest not only to government organizations in The Netherlands but also to all societies that value the enduring transparency and trustworthiness of their records.

C.6 Testbed infrastructure requirements

As explained in the main report, the testbed is intended to perform experiments to answer questions that arise in connection with digital records preservation projects to be carried out by the archives and by government agencies. Thus the testbed itself requires an infrastructure similar in many respects to the infrastructure required for actual (vs. experimental) digital preservation processes; in addition, it requires facilities for experimentation.

Below we briefly describe the resource requirements of the testbed. For ease of comparison, the discussion parallels the account of infrastructure requirements to support preservation processing (see Annex B, Section B.2).

C.6.1 Testbed processing requirements

In carrying out the seven iterative tasks just described, a testbed would perform most of the functions of an actual preservation process, but on a smaller scale and in an experimental setting. It would therefore need similar, although less intensive, computer processing capabilities. Such capabilities would include, at the start, ingesting digital records into the experimental environment and assessing their accessibility and usability on entry. Additionally, the testbed would need means for operating on the received material (e.g., copying and coding records, converting and generating metadata, searching for, retrieving and rendering located records, and so on).

Second, the testbed will need not only conversion and emulation tools but also will need to integrate them into a prototyping environment that enables development of experimental preservation procedures, development of validation tests, and so on. The flexibility required for prototyping (which is often inherently inefficient) implies a disproportionately greater need for processing power for the testbed than would be inferred from the relatively small number of records it will be required to process.

Third, the testbed will require a disproportionately wide range of types of processors (i.e., computing platforms) in order to evaluate the robustness of various ingestion, preservation, access, and rendering techniques. Proposals for creating the testbed infrastructure should be evaluated in part by the degree to which they address these multi-platform needs and the need for interoperability among them.

C.6.2 Testbed storage requirements

The testbed will need a repository for storing digital records and their metadata in varied forms, analogous to that of a real preservation system. However, the storage needs of the testbed are easier to estimate than those of the preservation process, and the storage demands of the testbed infrastructure are easier to meet than those of actual digital archives. For example, a testbed should enable a full range of experiments with varying amounts and types of metadata; but the relatively small number of records to be used in the testbed reduces the total amount of storage it will require. Further, size requirements will expand as a function of decisions made about the number and nature of documents selected for study rather than as a reflection of exogenous events.

Although metadata and data repository products are commercially available, many of these are rather strictly limited in how they can be used: such products should be carefully evaluated before concluding that they will satisfy the experimental needs of

the testbed. Generic, programmable database products may be preferable to more tailored systems based on rigid assumptions. Proposals for creating the testbed infrastructure should be evaluated in part by the degree to which they address this concern.

C.6.3 Testbed communications requirements

The communications requirements of the testbed, unlike those of a real preservation process, should be easily predictable and relatively modest. A local communications network will be adequate for most of its activities. It will probably be desirable to provide communication between the testbed and participating agencies who contribute experimental document samples not only to support ingestion processes but also to enable agency access to “preserved” records in the testbed repository (e.g., for validation tests). Available network connections should be more than adequate for such experimental purposes.

C.6.4 Testbed software requirements

The testbed will require the infrastructure to create prototype software to perform specialized functions of the preservation process, such as conversion, emulation and rendering. Even those capabilities that may ultimately be provided by commercial off-the-shelf software in the actual preservation process (such as record and metadata repository functions) are likely to require prototyping in the testbed, since one of the testbed’s purposes will be to develop specifications for the required functionality of such software.

The testbed will therefore require a flexible prototyping environment that allows experimental programs to be written and tried easily and effectively. Computer-aided software-engineering (CASE) tools and one or more prototyping languages or tools should be acquired for this purpose. CASE tools should be chosen at least partly on the basis of their ability to capture design rationales and decisions in reusable form. In addition, the CASE environment and programming tools chosen may be based on whatever existing expertise and experience resides within the National Archives (which is assumed to be the hosting organization for the testbed). Proposals for creating the testbed infrastructure should be evaluated in part by the degree to which they address this concern.

C.6.5 Testbed personnel requirements

Providing the testbed with the human resources required to carry out a successful program of research on the durability of digital documents will present a personnel challenge. Ideally, the testbed would include an individual who serves as the director of research, assuming responsibility for the overall substantive and methodological quality of its activities.

Additionally, the testbed would require a prototyping team comprising the following types of members: (1) Technical team members should include an architect or system designer, dedicated system support staff, software tool builders, prototype programmers, and a system programmer. (2) Archives and records management expertise must also be represented on the team, to help ensure that concepts and methods critical to preservation are appropriately implemented in digital media. They will, for instance, be key resources for carrying out testbed tasks 1 through 4 as well as task 6. (3) Representatives of record-originating agencies would, ideally, also be included on the team. Such individuals will be valuable at a number of points in experimental testbed processes (e.g., assisting document sample selection and ingestion; helping to corroborate results of validation tests with respect to documents from their agencies; and participating as experimental users in task 7). (4) Additional types of team members who could make valuable contributions to the testbed include: (i) representatives of software firms' R&D units knowledgeable about where relevant commercial offerings are headed; and (ii) members of the legal community whose expertise encompasses the interpretation of "evidentiality" in digital environments.

As an absolute minimum, two staff members with appropriate expertise within the National Archives should be dedicated to working on the testbed, in order to ensure continuity for the project and to begin to build intellectual and technical capital in this area within the Archives.

Proposals for creating the testbed infrastructure should be evaluated in part by the degree to which they offer on-site support for any proprietary hardware or software.

As suggested earlier, it would be desirable to include willing and interested agencies as affiliates or partners in any testbed effort. In part, such partnering will help to assure that experimental prototypes are designed with real world constraints in mind. In part, it will help to promote the diffusion and adoption of testbed results.

C.6.6 Testbed administration requirements

The testbed requires an administrative infrastructure that is a scaled down version of the administrative arrangements suggested for operational digital preservation processes. That is, it requires flexible agreements between the Archives (which is to host the testbed) and some number of agencies that are selected to work with the testbed, supplying sample records for experimentation. The full scope of the experiments to be undertaken in the testbed can be expected to evolve as preliminary results lead to further experiments, and the administration of the testbed should be flexible enough to accommodate this evolution.

Besides regular administrative arrangements, we also recommend the establishment of an advisory panel of relevant experts. The three domains of expertise needed for the testbed (described under personnel requirements in Section C.6.5, above) should be represented on the advisory panel. Along with constructive and critical feedback,

the panel might act as “judges” for validation tests and for the results of initial validation trials (see tasks 4 and 6). In addition, it would be desirable for the panel to include an interested and appropriate high-level government official to attract attention to and support for testbed efforts.

C.6.7 Testbed start-up requirements

The testbed infrastructure requirements described above are still somewhat generic. The following suggests a concrete “starter set” of resources that should be put in place in order to initiate the testbed. Additional requirements may become apparent as the testbed evolves, but the items listed here should allow testbed activities to get underway. The initial resources for the testbed should include the following.¹⁰⁶

- * 2 development class computers
 - Minimum 400 MHz single-processor
 - Minimum 128 MB RAM storage
 - Minimum 9GB fast-access (e.g., SCSI) disk
 - 10/100M Ethernet capability
 - Fast port (SCSI plus optional Firewire) plus serial/USB
 - Minimum 3 slots expansion capability
 - Minimum 19-inch high-resolution, color display
 - High-quality flatbed scanner with associated software
 - Graphics tablet (minimum A4 paper size) with associated software
 - Running Unix/X-windows environment or equivalent

- * 2 “workstation” computers
 - Representative of the highest-end workstations in use in the agencies chosen for initial involvement in the testbed
 - Minimum 300 MHz
 - Minimum 64 MB RAM storage

¹⁰⁶ Due to the rapid rate of evolution of computer hardware and software, the specific recommendations made here should be expected to remain valid for no more than six months to a year from the time of this writing (April, 1999). After that time, they should be extrapolated to produce equivalent recommendations at the time of use. For example, processor speeds and storage capacities should be expanded at the normal “inflation” rate for computer hardware, according to Moore’s Law (assuming it remains in force). Specific vendor and product names are avoided here, since these may also quickly become obsolete—and, more importantly, because such choices may need to depend on the results of open bidding processes. Exceptions to this strategy are made for specific products that are already present in the National Archives or with which resident staff already have experience.

- Minimum 4 GB disk
- Minimum 10M Ethernet capability
- Fast port (SCSI or Firewire), plus serial/USB
- Minimum 2 slots expansion capability
- Color display with size and resolution representative of displays in use in the agencies.
- Running whatever operating system is most representative of agency use (e.g., Windows, Windows/NT)

* 1 database server

(This may be overlapped with one of the development class machines, in which case its storage and software requirements should be added to those of the chosen machine.)

- Minimum 400 MHz single-processor
- Minimum 128 MB RAM storage
- Separate minimum 9 GB fast-access (e.g., SCSI) disk
- Running Unix/X-windows environment or equivalent
- Running a relational database management system (RDBMS)

* Internal/external network connections

- 10/100M Ethernet or equivalent internal, local area network (LAN) on all computers
- External Internet connection via LAN for all computers (or modem/dialup connection if and where LAN connection is infeasible)

* Software

- Relational database management system (RDBMS) and associated development tools from a major vendor (e.g., Oracle) for database server
- Entity-Relationship or equivalent data modeling tools for metadata design (if not packaged with DBMS) for both development systems
- Computer Aided Software Engineering (CASE) tools for both development systems (e.g., DDD, Designer/Developer/Discoverer)
- Programming languages, environments and support tool suites for both development systems (to be chosen based on staff experience and preferences)

- Scanning, OCR, and image-processing software to be determined by initial set of agency documents (for both development systems)
- SGML parsing/editing capability (FrameBuilder or equivalent) for metadata definition (for both development systems)

C.6.8 Testbed preservation metadata

The testbed requires a metadata framework for describing the digital records that are used in its preservation experiments.¹⁰⁷ Since the testbed is intended to be an experimental—rather than an operational—preservation environment, the metadata set presented for it here is not intended to represent a full archival metadata structure.¹⁰⁸ The metadata items included in this set consist of those needed to describe the technical aspects of digital preservation plus the minimum additional metadata necessary to enable the testbed to function as an experimental archival preservation environment.¹⁰⁹

Furthermore, the precise metadata requirements of the testbed may depend on the characteristics of the specific types of digital documents in the initial sample chosen for experimentation.¹¹⁰ It is expected that one of the results of the experiments performed using the testbed will be a better understanding of the kinds of technical

¹⁰⁷ Note that the term “metadata” as used in the recordkeeping context means data about records and their context. From a data modeling perspective, the proper term for such information is simply “data” rather than “metadata” since the latter term is reserved for information about the data in a database, such as descriptions of database fields in a data dictionary. We use the term “metadata” in the recordkeeping sense throughout this report, but data modelers must not allow this to confuse them in interpreting the data model and entity descriptions given in this section.

¹⁰⁸ General metadata considerations are discussed in Annex B, Section B.1.7. Assuming the testbed is successful, it should ultimately evolve into an experimental environment for testing operational concepts as well, in which case its metadata needs may expand to become those of a complete operational preservation environment. The framework presented here should therefore be implemented in an extensible manner so that it can evolve to meet such potential future requirements.

¹⁰⁹ For example, the testbed may require some general descriptive information about the records it uses in its experiments, as well as information about linkages between these records and others that may not be in the testbed itself. This kind of information would normally be part of a metadata-encapsulated record that is to be subjected to preservation, which would not be replicated in preservation metadata. However, the experimental records used in the testbed may come from disparate sources, which may make it difficult to guarantee the completeness or consistency of their attendant metadata. For this reason, the testbed metadata set includes a number of metadata fields that are not specifically related to preservation but are included simply to ensure that a standardized, consistent set of metadata will be available within the testbed. The source of each such field is shown as “Derived: from record metadata” in the following listing, and any of these that are found to be unnecessary in practice may be eliminated. In addition, some metadata may be unique to the testbed, providing information about the experimental context that would be irrelevant in an operational environment.

¹¹⁰ Similarly, the implementation details of the metadata elements proposed here may depend on the chosen database or prototyping environment and are left to the discretion of the National Archives, as are decisions about naming conventions, authoritative vocabularies, applicable standards for referencing and denoting sources, organizations, etc. (The published metadata frameworks cited in note 62 should be helpful in making such decisions.)

preservation metadata needed to characterize records. For example, whereas some published metadata schemes imply a relatively strong separation of technical preservation information from other functional or descriptive information about records, we hypothesize that such separation may not always be possible or desirable.¹¹¹ In order to avoid constraining the evolving metadata design of the testbed, we therefore caution against implementing (or even conceptualizing) it too rigidly.¹¹² The set of metadata elements proposed here is intended merely as a starting point.

The following metadata would provide a minimal capability for describing each digital document in the testbed sample. A notional entity-relationship data model for this testbed metadata repository is shown in Figure 7.¹¹³ The model consists of six entities, with unique identifiers shown as notional primary key attributes. In order to avoid cluttering the figure, non-key attributes (which would normally appear in the bottom half of each entity box) are replaced by descriptions of the contents of each entity (shown in parentheses). The actual non-key attributes of the entities are listed following the figure, grouped into what are here (and elsewhere) called elements and sub-elements.¹¹⁴ The EXPERIMENT entity and those elements and sub-elements of other entities that are required solely to support the testbed itself are shown in italics; items that are not italicized are representative of metadata that would appear in an operational preservation environment.¹¹⁵

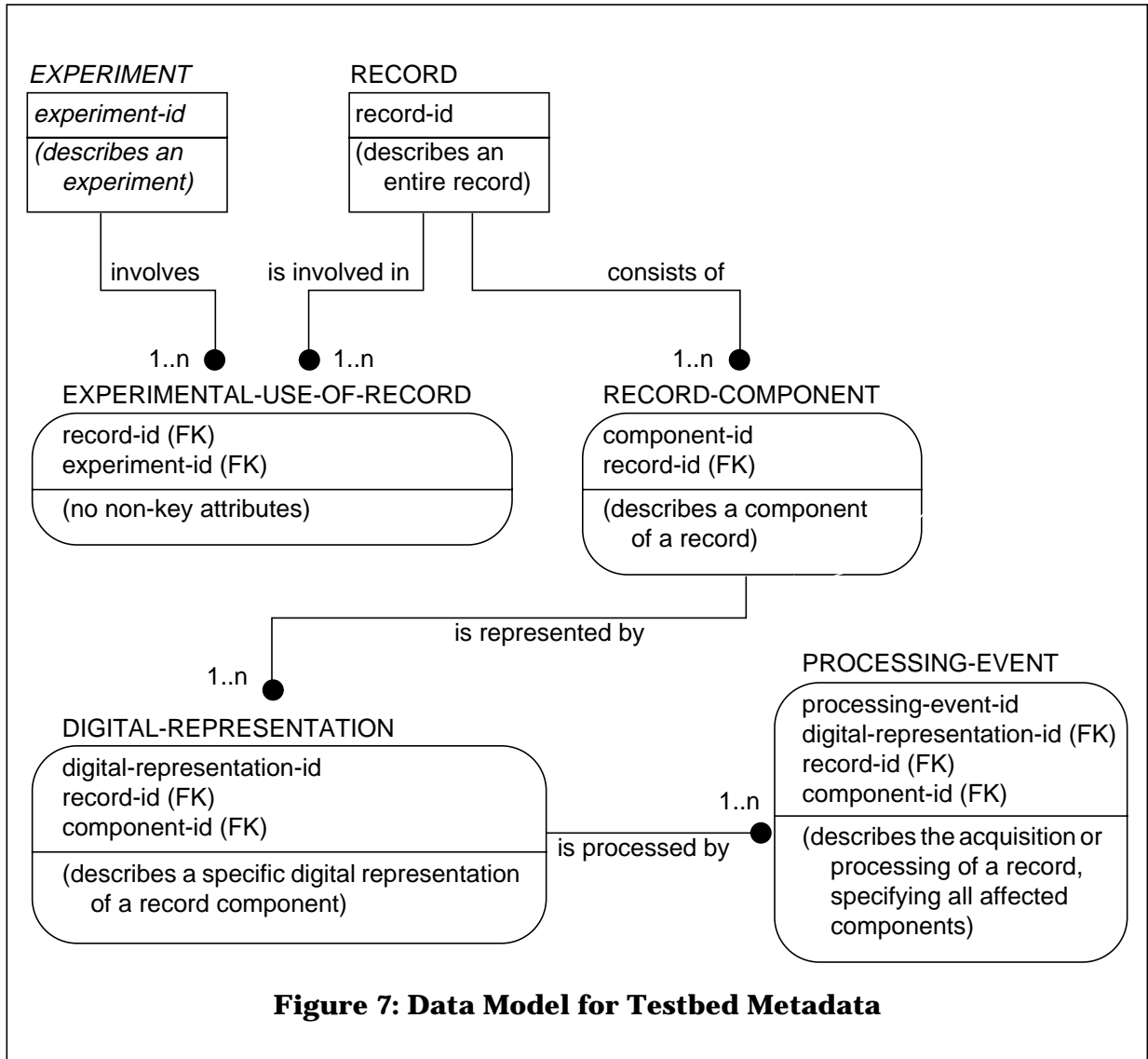
¹¹¹ For example, the Pittsburgh Functional Requirements reference model metadata (op cit, note 62) isolates technical information in its “Structural Layer” (which is something of a misnomer, since their model is not “layered” in the technical sense); on the other hand, the functional requirements of this reference model (available at <http://www.sis.pitt.edu/~nhprc/prog1.html>) do an admirable job of referencing relevant metadata elements wherever they apply in the functional model. The OAIS model (op cit, note 62) goes to the opposite extreme by including Representation Information along with a Data Object in its definition of an Information Object and then defining an open-ended set of possible Information Objects to represent Content Information, Preservation Description Information, Packaging Information, Descriptive Information, etc. Although this rightly packages representation information with every information object, it does not clarify (and may even obscure) the high-level relationship between technical preservation choices or constraints and the functional or behavioral capabilities of preserved digital records.

¹¹² The testbed metadata framework should be implemented using a flexible database environment or a programming environment with database features; for example, the widely available Prolog programming language supports relational database prototyping, which may be more flexible than using a formal database generation environment.

¹¹³ The figure uses IDEF1X graphics, except that the cardinality of relationships is shown numerically. All relationships are “defining” relationships of the form 1:1..n, as indicated. Independent entities are shown as rectangular boxes, whereas dependent entities have rounded corners.

¹¹⁴ Each of these elements and sub-elements is a “data element” in the database sense of that term.

¹¹⁵ Note that EXPERIMENTAL-USE-OF-RECORD is an “associative” entity, which requires no attributes of its own and consequently does not appear in the list of elements following the figure.



Entity: *EXPERIMENT*

EXPERIMENTAL CONTEXT DESCRIPTION

Testbed requirement: Mandatory (unique to testbed)

To explain the document's role in the testbed, intended experimental use, etc.

Form or Value set: Free text

Entity: RECORD

RECORD IDENTIFIER¹¹⁶

Testbed requirement: Mandatory

To facilitate experimentation and to be representative of an operational environment

Form or Value set: URI, ISBN, ISSN, X500, etc.

Derived: from record metadata (or created, if necessary)

TITLE

Testbed requirement: Mandatory

To be representative of an operational environment

Form or Value set: May contain alternative values (as sub-elements)

Derived: from record metadata

CONTENT DESCRIPTION/SUMMARY

Testbed requirement: Mandatory

To be representative of an operational environment

Form or Value set: May include text, images (e.g., thumbnails), etc.

Derived: from record metadata

AUTHENTICITY CRITERIA¹¹⁷

Testbed requirement: Mandatory

To define requirement for authentic preservation of the record

Sub-elements (optional):

Record type: Collection, Report, Letter, Memo, Weblet,¹¹⁸ etc.

Function description¹¹⁹

Relationships to other records (repeatable):¹²⁰

e.g., part/whole, version, reference, creative (i.e., "based on"), dependency

Behavioral description¹²¹

Form or Value set: Free text

Source: Created for testbed

¹¹⁶ This identifies the logical record as a whole, which may consist of multiple pieces, or components. Note that the term "resource" is often used in place of "record" in the names of metadata elements such as this.

¹¹⁷ This describes the criterion for deciding whether the record has been authentically preserved and recreated, as discussed in Annex A, Section A.1.2.

¹¹⁸ The term "weblet" denotes a related set of hypertext-referenced objects (or "web pages"), i.e., those that can be reached from some starting point by following hyperlinks satisfying some given criteria. See T. J. Watt, Jr., "Weblets: Fundamental Building Blocks for WWW Tool," *The Third International World-Wide Web Conference: Technology, Tools and Applications* (April 10-14, 1995, Darmstadt, Germany); www.igd.fhg.de/www/www95/proceedings/posters/55/index.html.

¹¹⁹ This may evolve into additional metadata elements, for example, representing information described in the Pittsburgh Functional Requirements (op cit, note 62) and/or formal business process models such as those currently under experimental development at the Dutch National Archives.

¹²⁰ Included if necessary to identify specific relationships that must be maintained for authenticity when they are not part of the ingested record metadata.

¹²¹ This must describe all aspects of the record's behavior that are relevant to its authenticity, including any aspects of its appearance, "look-and-feel" or interactive capabilities that must be recreated in order to understand the role it played as a record or to access, interpret or use it for archival or ongoing business process purposes (as discussed in Annex A, Section A.1.3).

SOURCE¹²²

Testbed requirement: Mandatory
To facilitate experimentation
as needed
Derived: from record metadata

AUTHOR/CREATOR (repeatable)

Testbed requirement: Optional
To be representative of an operational environment
Sub-elements: as needed
Derived: from record metadata

RIGHTS MANAGEMENT/ACCESS/USE CONSTRAINTS AND CONDITIONS¹²³

Testbed requirement: Mandatory
To facilitate experimentation with the preservation implications of such factors
Sub-elements:
Copyright
Redaction
Privacy
Freedom-of-Information
(Others as needed)
Derived: from record metadata

COMMENT (repeatable)

Testbed requirement: Optional
To facilitate experimentation
Sub-elements:
Comment proper
May include text, images (e.g., thumbnails), etc.
Name/title/role of person making this entry
Date-and-time of entry
Form or Value set: ISO8601 or equivalent
Source: Created for testbed

¹²² For example, this may specify the government agency that supplied the record.

¹²³ Though they may be outside the scope of initial testbed experiments, factors such as these may have implications for digital preservation, such as dealing with encryption, constraining the use of direct linkages to other records, etc.

Entity: RECORD-COMPONENT

LOGICAL TYPE OF RECORD/COMPONENT (repeatable)¹²⁴

Testbed requirement: Mandatory

To facilitate experimentation

Form or Value set: Physical, IMT (Internet Media Type), ISO standard, etc.

Source: Created for testbed

LOGICAL STRUCTURE OF RECORD/COMPONENT (repeatable)

Testbed requirement: Mandatory

To access logical digital records which may consist of multiple components

Sub-elements: as needed (technology-dependent)¹²⁵

Source: Created for testbed

LOGICAL DESCRIPTION OF COMPONENT (repeatable)¹²⁶

Testbed requirement: Mandatory

To facilitate access and use of logical digital components

Sub-elements:

Description of logical form and nature of component

Description of logical encoding (e.g., compression, encryption)

Note (repeatable)

(Others as needed: technology-dependent)

Source: Created for testbed

¹²⁴ A record will in general consist of multiple components, but components themselves may also be compound objects. Elements denoted as “record/component” are therefore used to describe records as a whole as well as their components and their sub-components, recursively. For an atomic (non-compound) record, there will be only a single instance of each such element.

¹²⁵ The value sets of sub-elements marked as “technology-dependent” will depend on the specific sample of digital document types chosen for the testbed and can be expected to evolve with the testbed. See also the Pittsburgh Functional Requirements reference model metadata (op cit, note 62), Structural Layer category III.E (Content Structure Metadata).

¹²⁶ Note a record may consist of a single component (described by one of these metadata elements), or it may consist of multiple components, each represented by one such element.

Entity: DIGITAL-REPRESENTATION

TECHNICAL DESCRIPTION OF COMPONENT (repeatable)¹²⁷

Testbed requirement: Mandatory

To document the original digital form of the record prior to ingestion into the testbed

Sub-elements:

Type (original, as-ingested, subsequent)

Explanation of how to access the component using this description

Applicable standards

Non-standard digital format, encoding, etc., if applicable

Documentation/references for interpreting the digital form of the record

Reference to intended application software¹²⁸

Intended application software documentation/references

Specific application software dependencies and implications

Reference to alternate acceptable software (e.g., “viewers”) and documentation

Environment software (e.g., operating system) dependencies and implications

Intended hardware characteristics

Display (resolution, aspect-ratio, color characteristics, etc.)

Storage (access/transfer speed and capacity ranges)

Intended range/type of input devices

(Others as needed)

Specific hardware dependencies and implications

Display characteristics (resolution, aspect-ratio, color characteristics, etc.)

Minimum expected processor/storage/display speeds

Maximum expected processor/storage/display speeds

Specific input device requirements

Additional processing dependencies (co-processors, graphics/sound cards, etc.)

(Others as needed)

Alternate acceptable hardware

Reference to acceptable physical or emulated hardware¹²⁹

Reference to documentation for physical or emulated hardware

Note (repeatable)

(Others as needed)

Source: Created for testbed

¹²⁷ This is intended to capture sufficient information about the digital format of each component of a record to facilitate preservation and use of the record as a whole; these elements are therefore expected to evolve as the testbed improves our understanding of the requirements for preserving digital records. Each subsequent form of an ingested record (including the ingestion form itself, if this differs in any way from the record’s original form) must be represented by its own instance of the DIGITAL-REPRESENTATION entity. Furthermore, generating each such form involves transforming the record from its previous form: this transformation must be documented by the CREATION/PROCESSING HISTORY elements of an instance of the PROCESSING-EVENT entity.

¹²⁸ Any sub-element marked as “reference to software” should point to a saved copy of the appropriate version of that software. Saved software must be accompanied by its own metadata specifying its source, version, hardware platform dependencies, and any ownership, lease, copyright, or other legal or operational constraints on its use.

¹²⁹ Any sub-element marked as “reference to hardware” should point to physical or emulated hardware that can run the required software.

Entity: PROCESSING-EVENT

CREATION/PROCESSING HISTORY (repeatable)¹³⁰

Testbed requirement: Mandatory

To facilitate experimentation

Sub-elements:

Name/title/role of person making this entry

Motivation/Rationale for processing performed

Component(s) affected¹³¹

Event/Process-type description

Occurrence-type (instantaneous, cyclic, occasional, etc.)

Comment (repeatable)

Date-and-time

Of this entry

Of event or initiation of process

Of completion of process (if applicable) or date of status being reported

Form or Value set: ISO8601 or equivalent

Description of Event or Process begun, completed, or in-progress

Overview of processing performed

Detailed description of processing performed

Reference to algorithms, programs, documentation, etc. (repeatable)

Name/title/role of person performing the processing described (repeatable)

Resulting Status of Record

Future projected processing requirements and schedule

Source: Created for testbed

COMMENT (repeatable)

Testbed requirement: Optional

To facilitate experimentation

Sub-elements:

Comment proper

May include text, images (e.g., thumbnails), etc.

Name/title/role of person making this entry

Date-and-time of entry

Form or Value set: ISO8601 or equivalent

Source: Created for testbed

¹³⁰ Instances of this element must document record ingestion as well any transformations or conversions of the record, whether performed for preservation purposes or other reasons.

¹³¹ A single instance of this element may be used to document the processing of individual components of a record or of the entire record, as appropriate.

Annex D: Additional technical background and context

This Annex presents additional technical background, context, and further justification of the scope and conclusions of the study.

D.1 Technical background and analysis

This section provides additional background and analysis of some of the technical issues related to digital preservation.

D.1.1 Digital recordkeeping and archiving

Documents, data, records, and informational and cultural artifacts of all kinds are rapidly being converted to digital form—if they are not created digitally to begin with. This rush to digitize is being driven by powerful incentives, including the ability to make perfect copies of digital artifacts, to publish them on a wide range of media, to distribute and disseminate them over networks, to reformat and convert them into alternate forms, to locate them, search their contents, and retrieve them, and to process them with automated and semi-automated tools. Yet the longevity of digital content is problematic for a number of complex and interrelated reasons.¹³²

As is now well appreciated, the physical lifetimes of digital storage media are often surprisingly short, requiring information to be “refreshed” by copying it onto new media with disturbing frequency. In addition, the technological obsolescence of these media (and of the hardware and software necessary to read them) pose an equally urgent threat. Moreover, most digital documents and artifacts require software to bring their bit streams to life and make them truly usable; as these programs (or the hardware and software environments on top of which they run) become obsolete, the digital documents that depend on them become unreadable—held hostage to their own encoding. As pointed out in the body of this report, this problem is particularly ironic, since digital documents can be copied perfectly, which at first glance appears to make them eternal.

There is currently no demonstrably viable technical solution to the problem of ensuring the longevity of digital information; yet if it is not solved, governmental records, which are becoming increasingly digital, are in grave risk of being lost.

¹³² See ACCIS (Advisory Committee for the Coordination of Information Systems). *Management of Electronic Records: Issues and Guidelines*, New York: United Nations, 1990; Lesk, Michael. *Preserving Digital Objects: Recurrent Needs and Challenges*, <http://community.bellcore.com/lesk/auspres/aus.html>; Morris, R. J. “Electronic Documents and the History of the Late Twentieth Century: Black Holes or Warehouses,” in *History and Electronic Artefacts*, Edward Higgs, ed., Oxford: Clarendon Press, 1998, pp. 31-48; Popkin, J. and A. Cushman. *Integrated Document Management—Controlling a Rising Tide*, Stamford CT: Gartner Group, 1993; Rothenberg, 1995a, op cit, note 96; *Time & Bits: Managing Digital Continuity*, <http://www.ahip.getty.edu/timeandbits/intro.html>.

In addition to the technical aspects of this problem, there are administrative, procedural, organizational, and policy issues surrounding the management of digital records. Numerous features of digital records distinguish them from traditional paper records in ways that have significant implications for how they are generated, captured, transmitted, stored, maintained, accessed, and managed. Paramount among these differences is the greatly reduced lifetime of digital records without some form of active preservation: this mandates new approaches to saving digital records to avoid their loss. Non-technical issues include questions of jurisdiction, responsibility for various phases of the existence of digital records, funding, and policies requiring government agencies to adhere to standard techniques and practices to prevent loss of digital information. Nevertheless, it is difficult to address these non-technical issues meaningfully in the absence of a defined technical solution to the digital longevity problem.

D.1.2 Technical dimensions of the digital preservation problem

Although the preservation and management of digital records involves interrelated technical, administrative, procedural, organizational, and policy issues, a sound technical approach forms the foundation on which everything else must rest.¹³³ Preserving digital records may require substantial new investments and commitments by institutions and government agencies, requiring new economic and administrative policies for funding and managing digital preservation, but it is impossible to allocate responsibilities or assess costs for an undefined process.¹³⁴ Until a viable technical approach to digital longevity has been identified, developed, and proven viable, it is premature to spend too much effort designing the administrative and organizational environment that will embed this approach.¹³⁵ This section attempts to identify the technical dimensions of the problem.

D.1.2.1 Digital media suffer from physical decay and obsolescence

There is generally widespread awareness of the fact that digital storage media have severely limited physical lifetimes, often on the order of just a few years. Moreover, even if archival quality media were developed, they would probably fail in the market, since they would quickly be made obsolete despite their physical longevity by newer media having increased capacity, higher speed, greater convenience, and lower price. No matter how long a disk, tape, etc. may last, it is unlikely to remain readable for more than a few years, because its form, size, and recording format are likely to become obsolete within that time.¹³⁶

¹³³ The details of a technical solution may themselves depend on the administrative, organizational, and policy environment in which that solution is implemented—the technical and administrative aspects of this problem are in fact interrelated in complex ways—yet a sound technical approach must be identified before significant progress can be made.

¹³⁴ It is meaningless to ask which organizations should perform which aspects of the digital preservation process or what procedures they should follow without having first identified a technically viable process.

The dual problems of media lifetime and media obsolescence have led to the recognition that digital information must be copied to new media (“migrated” or “refreshed”) on a very short cycle (i.e., every few years). Copying is a straightforward solution to these media problems, but it is not trivial: the copy process must avoid corrupting information via compression, encryption, or changing data formats.

In addition, as media become more dense, each copy cycle aggregates many disks, tapes, etc. onto a single new unit of storage (say CD or DVD). This raises the question of how to retain any labeling information that may have been associated with the original media: since it is infeasible to squeeze the contents of the labels of 400 floppy disks to fit on the label of a single CD, label information must be digitized to ensure that it adheres to the records it describes. But whereas labels are directly human readable, digitized information is not: labels must therefore be digitized in such a way as to remain more easily readable by humans than the records they describe.¹³⁷

D.1.2.2 Digital records depend on software

Yet media problems are merely the tip of the iceberg. Far more problematic is the fact that digital records are in general dependent on application software to make them accessible and meaningful. Copying media at most ensures that the original bit stream of a digital document will be preserved. But a stream of bits cannot be made self-explanatory, any more than hieroglyphics were self-explanatory before the discovery of the Rosetta Stone. A bit stream may encode text, data, imagery, audio, video, animated graphics, and any other form or format, current or future, singly or

¹³⁵ Though few research efforts in the international archival community have focused on the technical aspects of long-term preservation of digital records, the ongoing work of the InterPARES project includes a number of groups (in addition the Dutch National Archives itself) whose work is relevant to long-term preservation. John McDonald, Luciana Duranti, and Charles Dollar in Canada have published widely on theoretical aspects of archival recordkeeping (McDonald 1998, op cit, note 77; C. M. Dollar. *Archival Theory and Information Technologies: The Impact of Information Technologies on Archival Principles and Practices, Information and Documentation, Series #1*, Oddo Bucci, ed., Macerata, Italy: University of Macerata, 1992.), while Sue McKemmish, Frank Upward, and others in Australia have espoused the idea of the “records continuum” (McKemmish and Parer, 1998, op cit, note 3, and *Structuring the Records Continuum, Part Two*) by Frank Upward, 1998, Monash University Records Continuum Research Group).

In addition, Ivar Fønnes and others at the Norwegian National Archives have developed the Noark approach to using standards for archival interchange (Fønnes, 1998, op cit, note 80), while other work in Scandinavia is attempting to produce logical descriptions of the behavior of relational databases that would allow encoding records from such databases as text (see Bikson and Frinking, 1993, op cit, note 74).

Related efforts in the U.S. are described in note 62. Also relevant is the U.S. Department of Defense’s *Design Criteria Standard for Electronic Records Management Software Applications* (widely though erroneously referred to as the “DoD Guidelines”), DoD 5015.2-STD, which can be found at <http://jttc.fhu.disa.mil/rec/mgmt/>. Though it provides some general requirements for designing records management applications, this has relatively little to say on the subject of preservation.

¹³⁶ See K. Schurer, “The Implications of Information Technology for the Future Study of History,” in *History and Electronic Artefacts*, Edward Higgs, ed., Oxford: Clarendon Press, 1998, pp. 155-168.

¹³⁷ See D. Bearman, “Documenting Documentation,” *Archivaria*, Vol. 34 (Summer), 1992.

combined in a hypermedia lattice of pointers. In general, a bit stream can be made intelligible only by running the software that created it or some closely related software that understands it.

This implies that digital records exist only by virtue of software that understands how to access them; they come into existence only by running this software. Therefore, the only reliable way (and often the only possible way) to access the meaning and functionality of a digital document is to run its original software (or some closely related software that understands it).¹³⁸ Yet such application software becomes obsolete just as fast as the digital storage media discussed above. And although we can save obsolete programs (and the operating system environments they require), running them requires computer hardware, which becomes obsolete just as quickly. It is therefore not trivial to use a digital document's original software to view the document on some unknown future computer. This is the crux of the technical problem of preserving digital documents and is the target of the emulation solution discussed below.

D.1.2.3 Additional considerations

Any technical solution must also be able to cope with issues of corruption of information, privacy, authentication, validation, and preserving intellectual property rights. This last issue is especially complex for records that are “born digital” and therefore have no single “original” instance, since traditional notions of copies are inapplicable to such records. Finally, any technical solution must be feasible in terms of the societal and institutional responsibilities and costs required to implement it.

D.1.3 Criteria for an ideal solution to digital preservation

An ideal approach to digital preservation would provide a single, extensible, long-term solution that can be designed once and for all and applied uniformly, automatically and in synchrony to all types of documents and all media with minimal human intervention. It should provide maximum leverage, in the sense that implementing it for any document type should make it usable for all document types. It should facilitate records management (cataloging, disposal, etc.) by associating human readable “labeling” information and metadata with each document. It should retain as much as desired (and feasible) of the original functionality, look, and feel of each original document while minimizing translation (to minimize both labor and the potential for loss via corruption). If translation is unavoidable (e.g., when translating labeling information or metadata) this should be guaranteed to be reversible, so that the original form can be recovered without loss.

¹³⁸ D. Swade, “Preserving Software in an Object-Centred Culture,” in *History and Electronic Artefacts*, Edward Higgs, ed., Oxford: Clarendon Press, 1998, pp. 195-206.

An ideal approach should offer alternatives for levels of safety and quality, volume of storage, ease of access, etc. at varying costs, and it should allow changing these alternatives for a given document, document type, or corpus at any time in the future. It should provide single-step access to all documents and should offer up-front “acceptance testing” at accession time, to demonstrate that a given document will be accessible in the future. Finally, the only assumptions it should make about future computers are that they will be able to perform any computable function and (optionally) that they will be faster and/or cheaper to use than current computers.

D.1.4 Analysis of existing and previously proposed approaches

A number of approaches have been proposed or are beginning to come into use in an attempt to achieve digital longevity, but none of these can yet be said to have been proven as a viable long-term digital preservation approach. Some rely on well-known computer science techniques but have not yet been shown to be capable of preserving authentic digital records in the sense discussed in this report. Others seem to have great potential but rely on speculative techniques or speculative adaptations of existing techniques. Furthermore, since the oldest digital records are still relatively young, none of these proposed techniques has yet been tried and evaluated on more than a small sample of digital records over a short time span.¹³⁹

It is generally assumed that digital records will ultimately be preserved by using a number of different approaches (such as those discussed here). This assumption is based on the reasonable conjecture that no single solution will offer a combination of capability, cost, and pragmatic factors that will be optimal for all classes of digital records and all archival settings. While this appears to be a safe (i.e., conservative) assumption, it stands in need of empirical validation. In particular, the economy of scale of using a single, universal preservation approach (if one could be devised) might well outweigh the advantages of using alternative, specialized preservation approaches for different situations. Since the digital preservation problem arises at least in part from the proliferation of incompatible digital representations in the first place, it seems counterproductive to intentionally proliferate digital preservation approaches without at least attempting to find a single approach that can preserve all (or at least most) digital documents in a uniform way. While it is of course important to allow for the possibility that multiple approaches will be required, it seems strategically preferable to seek solutions that offer some degree of universality (as well as simplicity and cost-effectiveness).

With this in mind, we advocate an open-minded, experimental investigation of alternative preservation approaches, as embodied in our preservation strategy (presented in Annex A) and our testbed design (presented in Annex C). The following

¹³⁹ In addition, since the forms of digital records are still evolving rapidly, it is difficult to guarantee that current experiments or demonstrations will generalize to future types of digital records over archival timescales.

discussion presents additional technical analysis that was performed as part of this study. For each proposed preservation approach, we analyze technical, procedural, and organizational issues that we have identified.¹⁴⁰

D.1.4.1 Reliance on printing

It is sometimes suggested that digital records be printed (or engraved microscopically) and saved in “hard” copy. This is not a true solution to the problem, since many documents (especially those that are inherently digital, such as hypermedia) cannot meaningfully be printed at all or would lose many of their uniquely digital attributes and capabilities if they were printed. Even digital renditions of traditional documents (such as linear text) lose their “core digital attributes” by being printed; that is, they sacrifice being directly machine-readable, which means they can no longer be copied perfectly, transmitted digitally, searched or processed by computer programs, etc. Similarly, attempting to save digital documents by printing the 0s and 1s of their bit streams on paper (or engraving them in metal) sacrifices their machine-readability and the core digital attributes that it enables.¹⁴¹

Moreover, printing destroys any interactive or dynamic functionality that a digital document may have, since the document’s original software can no longer be run on it. For all of these reasons, saving digital documents by printing them is not considered to offer a solution to the core of the problem of digital preservation.

¹⁴⁰ In addition, see Rothenberg, J., *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation: A Report to the Council on Library & Information Resources (CLIR)*, January 1999, which can be read at <http://www.clir.org/pubs/reports/rothenberg/contents.html> or downloaded from <http://www.clir.org/pubs/reports/rothenberg/pub77.pdf>.

¹⁴¹ If a printed bit stream could be read back into a computer with minimal likelihood of error, then this would simply be an alternative form of digital storage. The fact that the 0s and 1s of a printed bit stream could be read by humans without additional mechanism (except perhaps for a microscope) would not offer a significant advantage over most digital storage media, because a digital document saved in this way could still not be understood by a human reader without the use of software. That is, printing eliminates a document’s reliance on software only if the document can not only be read but can also be directly understood by a human reader; yet a printed bit stream would still require interpretation by software before it could be understood by a human. Printing bit streams in human-readable form would therefore not make them understandable without the use of software, and (at least using current technology) it would make them harder for a computer to read with a low enough error rate for them to qualify as machine-readable, thereby making it harder to read them with the software needed to make them understandable.

For discussions of the primacy of digital records over printed surrogates, see D. Bearman, “The Implications of *Armstrong v. Executive Office of the President for the Archival Management of Electronic Records*,” *American Archivist*, Vol. 56, 1993, pp.150-160. and United States District Court for the District of Columbia: Opinion of Charles R. Richey, United States District Judge, January 6, 1993.

D.1.4.2 Reliance on standards

On the face of it, standards appear to offer a solution by representing digital documents in forms that will endure into the future and for which future software will always provide accessibility. For example, the relational database (RDB) has been offered as a paradigmatic example of how this might work.¹⁴² It is argued that since all relational database management systems (RDBMSs) are based on the same mathematical foundation,¹⁴³ any RDB can in principle be translated without loss into the specific form of RDB used by an archives. Even if the RDBMS used by the archives is later changed, all of the RDBs should be able to migrate to the new RDBMS without loss, since all RDBMSs support the same functionality.¹⁴⁴

While this argument appears convincing, it fails in several significant and revealing ways. First, though the relational model mandates a standard baseline of functionality, real RDBMSs distinguish themselves in the marketplace by introducing proprietary features that extend the relational model (such as such as unique diagrams for data modeling, outer joins, support for views, business rules, triggers, etc.) Any RDB that uses such features becomes at least somewhat non-standard and will lose some of its functionality if translated into some other RDBMS.¹⁴⁵ Users are motivated to use such features because they provide additional functionality, but using them creates unique, non-standard documents.

Furthermore, far from being a representative example, the relational database is actually unique: no other kind of digital document rests on a formal mathematical foundation that can serve as the basis for its standardization. Most standards are informal, ad hoc, and often short-lived; and since they lack a formal underpinning, they often compete with rival standards, leading to the well-known joke: “The best thing about standards is that there are so many different ones to choose from!”¹⁴⁶

Finally, just as the relational paradigm replaced earlier network and hierarchical database paradigms, it is itself now under attack by the new object-oriented database (OODB) paradigm, which may eventually replace it. And as was the case with previous database paradigm shifts, the transition from relational to object-oriented database cannot be performed by automatically translating RDBs into OODBs. The

¹⁴² See NARA, *Management, Preservation and Access For Electronic Records with Enduring Value*, July 1, 1991, and K. Thibodeau, “To Be Or Not to Be: Archives for Electronic Records,” in *Archival Management of Electronic Records*, Archives and Museum Informatics Technical Report No. 13, ISSN 1042-1459, D. Bearman, ed., 1991.

¹⁴³ E. F. Codd, “Relational Database: A Practical Foundation for Productivity,” *CACM*, Vol. 25, No. 2, February 1982, pp. 109-117.

¹⁴⁴ See ACCIS, 1990, op cit, note 132 and ACCIS. *Strategic Issues for Electronic Records Management: Toward Open System Interconnection*, New York: United Nations (Advisory Committee for the Coordination of Information Systems), 1992.

¹⁴⁵ See Bikson and Frinking, 1993, op cit, note 74.

¹⁴⁶ This classic remark is generally attributed to Andrew Tanenbaum.

paradigms are so different that such translation is typically meaningless: even if it is possible, the result may well possess neither the formality of the original relational form nor the semantic expressiveness of the new object-oriented form. Even the best standards are often bypassed and made irrelevant in this way by the inevitable paradigm shifts that continue to characterize information science.

It is often argued that the solution to paradigm shifts is to force digital documents into current standard forms (even if this loses some of their functionality) and then translate them when current standards become obsolete into whatever standards supplant the obsolete ones.¹⁴⁷ This is analogous to translating Homer into modern Dutch by way of every intervening language that has existed during the past 2500 years, which would undoubtedly produce an unrecognizable result. Translation always loses something and rarely allows us to recover the original by translating backwards again.

Despite all this, standards should not be dismissed. Some standards (such as SGML and its offspring, such as XML) have proven highly extensible and worthwhile within their limited scope.¹⁴⁸ As information science continues to define itself, new long-lived standards may eventually emerge that will be capable of preserving authentic records of various kinds; these should be evaluated as they appear. In the meantime, the judicious use of standards may afford some breathing space by allowing digital records of certain kinds to be preserved over short or medium time scales, while awaiting longer-term solutions.¹⁴⁹

From a preservation perspective, the main shortcomings of standards are: (1) they do not—and are not in the foreseeable future likely to—encompass all forms of digital records, so that relying on them will force some records to be either abandoned entirely or corrupted by being translated into standard forms that do not do them

¹⁴⁷ This is closely related to migration, as discussed in Section D.1.4.4 below.

¹⁴⁸ In fact, if SGML had been adopted as a common inter-lingua among word processing programs, it would have greatly relieved the daily conversion problems that plague most computer users. That this has not occurred is symptomatic of the fact that even good standards do not necessarily sweep the marketplace (see Bikson, 1997, *op cit*, note 81).

¹⁴⁹ A few archives have developed preservation approaches based on the use of standards as a way of eliminating the need to handle digital records in arbitrary formats. For example, the Norwegian National Archives (see Fonnes, *op cit*, note 135) has expanded their Noark system to encompass digital records, defining four acceptable formats (ASCII, SGML/XML, TIFF and PDF) for the exchange and transfer of such records across the records continuum. Part of their strategy has been to encourage the development of Noark-compliant application programs that perform various recordkeeping functions with records in these formats. This gives agencies an incentive to use the Noark standard forms for their digital records, since doing so allows them to use these Noark applications. Note that while the particular formats included under the Noark umbrella are eminently reasonable choices at the moment, none of them is likely to survive for very long. ASCII will probably give way to Unicode in the near future, XML is still evolving, and PDF is (so far, at least) a proprietary product of Adobe Software. While it may be tempting to believe that the very ubiquity of these standards will force the market to create well-defined and automated migration paths from these formats into their successors, this may be yet another example of the kind of “wishful thinking” that characterizes migration, as discussed in Section D.1.4.4 below.

justice; and (2) since they do not last indefinitely, the use of standards must be combined with migration, which entails the risk that records will inevitably be corrupted as they are converted into new forms, as well as the other disadvantages of migration discussed below.

In summary, converting digital documents into standard forms and migrating to new standards if necessary may be a useful interim approach while a true long-term solution is being developed. In addition, standards may play a useful role in a long-term solution by providing a way to keep metadata and annotations readable. However, standards do not by themselves appear to offer a true solution to the problem of long-term digital preservation.

D.1.4.3 Reliance on computer museums

To avoid the dual problems of corruption via translation and abandonment at paradigm shifts, some have suggested establishing computer museums where old machines would run original software to access obsolete documents.¹⁵⁰ While this carries a certain degree of technological charm, it is flawed in a number of fundamental ways. It is unlikely that old machines could be kept running indefinitely at any reasonable cost, and even if they were, this would limit access to old digital documents to a very few sites in the world, thereby again sacrificing many of these documents' core digital attributes.

Furthermore, this approach ignores the fact that old digital documents (and the original software needed to access them) will rarely survive on their original digital media, for reasons discussed above. If an obsolete digital document and its software survive into the future, it will probably be because their bit streams have been copied onto new media that did not exist when the document's original computer was current. The document would therefore have to be read by an obsolete machine from a new medium for which that machine has no physical drive, interface, or device software. This would require building unique new device interfaces between every new medium and every obsolete computer in the museum as new storage media evolve, as well as coding device drivers for these devices, requiring the maintenance of systems programming skills for each obsolete machine. This seems hopeless.

Finally, computer chips themselves have limited physical lifetimes, decaying by means of mechanisms such as metal migration and dopant diffusion. Even if they were stored carefully, maintained religiously, and never used, such mechanisms would eventually render obsolete computers inoperative. For all of these reasons, computer museums do not offer a serious solution.

¹⁵⁰ See Swade, 1998, op cit, note 138.

D.1.4.4 Reliance on migration

The approach that most institutions are adopting (if only by default) is to expect digital documents to become unreadable or inaccessible as their original software becomes obsolete and to translate them into new forms as needed whenever this occurs.¹⁵¹ This is the traditional computer science “migration” approach, adapted to the preservation of digital records.¹⁵²

Migration is a well-known technique that has been used for decades to convert old data and documents into new formats.¹⁵³ Its attraction as a preservation approach is that, given enough effort and time, it should often be possible to convert records into some new form that will allow them to be accessed by new software. Depending on the applicable authenticity criteria, this conversion may or may not preserve enough of a record’s attributes to constitute true preservation, but conversion is certainly preferable to standing by and watching as records become inaccessible.

The problem with migration is that it involves conversion, which is generally both expensive and risky (in that converted documents may be corrupted in a wide variety of ways). Conversion is labor-intensive and costly, since each different document type and each different transition from an obsolete set of software tools to a new one requires the development of a specialized, often unique conversion strategy. The actual conversion process produced by this strategy may vary from relatively trivial (for example, performing a simple remapping of character codes—though even this is rarely as straightforward as it sounds) to arbitrarily complex (for example, redesigning a relational database to be object-oriented). Paradigm shifts, such as the change from relational to object-oriented database, represent the worst cases, since they are (by definition) impossible to predict and require conversions that can rarely be performed without considerable human intervention (as well as loss).

In addition to its cost, conversion runs the risk of corrupting records, since all but the most trivial translation processes change meaning in subtle or significant ways. Legal restrictions may explicitly prohibit “reformatting” digital records in this way and may not recognize converted records as equivalent to their original forms. The potential for corrupting a record in a single conversion step is compounded by the need for successive conversions if migration is used for long-term preservation.¹⁵⁴

¹⁵¹ See Bikson and Frinking, 1993, op cit, note 74 and Dollar, 1992, op cit, note 135.

¹⁵² Though this approach is often linked to the use of standards, standards are not intrinsically a part of migration.

¹⁵³ “Migration” is also used to denote the process of revising or upgrading software so that it can run in new software environments or on new hardware. This may sometimes play a role in document migration, since it may be desirable to convert the software required to read a given type of document as well as converting the documents themselves. Furthermore, in the case of documents that are themselves programs, document migration will mean software migration.

¹⁵⁴ See also note 17.

The accumulation and compounding of corruption by successive translations is analogous to the propagation of round-off error in numerical analysis or the parlor game in which a sentence is whispered from one person to the next, typically being changed each time until it becomes a parody of the original. This potential for corruption is inherent in the process of successive conversion: even if each conversion is performed with great care, subtle differences between successive formats and formalisms may result in the inevitable loss of certain attributes of the original digital document, which (again, depending on the applicable authenticity criteria) may result in the loss of authenticity or meaning. To reiterate, this risk does not derive merely from the danger of inept application of the conversion process (though that increases the risk): it is an inescapable aspect of conversion itself. Furthermore, when paradigm shifts occur, meaningful conversion may be virtually impossible, in which case records may be abandoned entirely or corrupted beyond recognition.

From a preservation perspective, migration is essentially an approach based on wishful thinking. Since we cannot know how things will change in the future, we cannot predict what we will have to do to keep a given digital document (or type of document) accessible and readable in the future. We can merely predict that we will often have to do something, since software and hardware will become obsolete and paradigms will shift in unpredictable ways. Since paradigm shifts apply to different kinds of documents and records, we must expect such shifts to force the migration of different kinds of documents on independent, unrelated schedules, with each document type (and perhaps even individual documents or their components) requiring specialized, labor-intensive handling.

Because of these inherently unknowable factors, we cannot predict how much effort, time, or expense will be required to perform any given migration cycle, how frequently migrations will have to occur, how successful we will be in each case, how much we will lose in each translation, or how many records will be corrupted or lost in each cycle. Nor can we expect each cycle to derive much benefit from previous cycles, since each migration will pose a new set of unique problems. Furthermore, a successful demonstration of migration on a particular collection of digital records proves little about the future feasibility of the approach, since the migration of a given document type into successive new paradigms (not to mention the migration of future document types) may require unique, specialized processing.

Finally, whenever a given type of digital document threatens to become obsolete, conversion must be applied to every document of that type before the old software on which those documents depend—and the expertise required to perform the conversion appropriately—becomes unavailable. That is, it is not only expensive to develop each conversion strategy required to perform migration and often expensive to perform this conversion on a single document, but *every* document that is to be preserved must be converted, which may involve a huge cost at each migration cycle. Though it may be rare for more than a few percent of the records in a given archival collection to be accessed in the period of time that can be expected to occur between migration cycles, *all* such archival records will have to be converted, since it cannot

be known which ones will need to be accessed—and any records that are not converted in a given migration cycle may be lost forever. Migration cannot be performed “on demand” for a given record when it is actually accessed but must be performed proactively for all records, since failing to do so makes them inaccessible. Migration therefore incurs a potentially high per-record conversion cost multiplied by a huge (and presumably increasing) number of records that must be preserved each time migration occurs; and this high resultant cost must be multiplied by the indefinitely large number of migration cycles that will recur unendingly throughout the future.

D.1.4.5 Reliance on “viewers”

In a variant of migration, a number of organizations have advocated relying on generic programs called “viewers” that can interpret and display the results of a range of current formats, such as appear in word processing or graphics documents. This may avoid having to collect and maintain a large software library containing all current application programs that might be needed to view a given range of formats, instead choosing a few carefully selected viewers that can do the same.

This appears to be a cost-effective technique, though it must be demonstrated that a given viewer recreates all of the attributes that each digital record exhibited when viewed using the original software associated with that record (or at least all of those attributes that are relevant to the given authenticity criteria).

However, from a preservation perspective, this approach is essentially the same as using any application software to view specific digital formats. Eventually, any such software becomes obsolete, either because it fails to correctly interpret and view new versions of the same digital formats or because it ceases to run in a modified software environment or on new hardware. In fact, the use of a viewer may make an organization all the more dependent on a single piece of software for accessing many different kinds of records, and all the more vulnerable to bugs in that single program, lack of support for it, or its becoming obsolete. When this happens, either the records themselves must be converted to a new form, accessible by some new viewer, a new viewer must be acquired, or the viewer must migrate to a new computing environment. The latter option requires rewriting the viewer program, which assumes that the organization either has the rights to the source code for the viewer or is able to “reverse engineer” the running program (a laborious process that has been described as analogous to trying to recover a pig from sausage).

D.1.4.6 Reliance on “digital archaeology”

An interesting variation on the migration theme is to perform it only when necessary, in the future, by means of what is sometimes informally referred to as “digital archaeology” (suggesting that old bits can be dug up and interpreted). The idea here is to save the original bit stream of a digital document without performing any conversion or other processing to preserve its meaning until it is desired to read it in

the future. At that time, the bit stream is analyzed, perhaps with the help of software that knows something about obsolete digital formats, in an attempt to convert it into some readable, understandable form. This can be thought of as delayed (or “lazy”) migration.¹⁵⁵

Digital archaeology runs the risk of losing documents, since by the time they are required, it may be impossible to decode their bit streams. However, it has the great advantage of avoiding the expenditure of time, effort and money required to convert all digital documents (the vast majority of which may never be required to be read at all) whenever their formats become obsolete. Furthermore, this approach is amenable to a range of implementations in which varying degrees of support are provided for the archaeological process. For example, the more information (i.e., metadata) is saved about the formats of various document types or individual documents, the easier it will be to decode them if necessary in the future. If the bit streams of the programs used to read obsolete documents are also saved—even without any scheme for allowing those programs to be run on future computers—they may also be decoded in the future (either “by hand” or using future programs that understand obsolete programs) as an aid to decoding the documents they once accessed.

This approach is based on wishful thinking to an even greater degree than migration, since it simply assumes—without being able to demonstrate—that it will be possible to decode arbitrary obsolete digital documents as needed in the future, provided that we have saved their original bit streams. On the other hand, it is the least costly approach (except for not saving digital documents at all) since it requires nothing but saving the bit streams of documents.¹⁵⁶ Furthermore, this approach can be thought of as a backup or emergency recovery technique that is compatible with any other approach that saves the original bit streams of digital documents (which migration does not) since it can be applied if the primary preservation technique fails for some reason.

D.1.4.7 Reliance on saving the bits

A final approach that has been suggested is to save the original bit streams of digital records and somehow create metadata that explains how to interpret those bit streams. This is related to the digital archaeology approach described above, but it

¹⁵⁵ The term “lazy” as used in computer science has no negative connotation: it means performing some process only when its results are required, rather than in advance, thereby avoiding having to perform it at all if its results are never needed. The term “just-in-time” would be misleading in the case of delayed migration, since waiting until a document is needed may be too late (i.e., *not* in time) to allow successful interpretation. Note that if a document is translated into a form that is directly understandable to humans (for example by extracting its textual content), it may be argued that this would not be an instance of migration at all, since it would not result in converting the document into some new digital format; we therefore prefer the term digital archaeology to delayed or lazy migration.

¹⁵⁶ We reiterate, however, that saving bit streams is not trivial, since it requires refreshing media and ensuring that the bit streams are copied without corruption.

attempts to preserve bit streams by creating metadata that will enable decoding them in the future. The Universal Preservation Format (UPF) developed by WGBH Public Broadcasting in Boston (based partly on earlier work at Apple and elsewhere on the Bento standard) advocates this kind of approach for digital music and video recordings,¹⁵⁷ while the Digital Rosetta Stone¹⁵⁸ advocates it for general digital material.

These approaches recognize the fact that migration, with its repeated conversion, will inevitably corrupt a digital record beyond recognition. However, they have yet to demonstrate how to encode the behavior of arbitrarily complex software into metadata to allow the interpretation of saved bit streams. Whereas it may be possible to provide mathematical descriptions for the interpretation of simple digital encodings such as recorded music, encoding the behavior of arbitrary software in this way remains an unsolved problem in computer science.

D.1.4.8 Reliance on emulation

In light of the foregoing analysis, we hypothesize that the best (if not the only) way to satisfy the ideal criteria for a preservation approach (suggested in Section D.1.3) is to somehow run a digital record's original software. This appears to be the only reliable way of recreating a digital record's original functionality, look, and feel. The approach proposed here is to enable the emulation of obsolete systems on future, unknown systems, to allow running a digital record's original software in the future despite its being obsolete. Though it may not be feasible to preserve every conceivable attribute of a digital record in this way, it should allow us to recreate the record's behavior as accurately as desired—and to test this accuracy in advance.¹⁵⁹

The next three sub-sections describe this approach in considerable detail, since it has not been given nearly as much attention as migration in the literature.¹⁶⁰

¹⁵⁷ See *The Universal Preservation Format*, Dave MacCarn and Thom Shepard, WGBH Educational Foundation, Draft Revision December, 1998.

¹⁵⁸ *Digital Rosetta Stone: A Conceptual Model for Maintaining Long-term Access to Digital Documents*, Alan R. Heminger, and Steven B. Robertson, 1996(?).

¹⁵⁹ To the extent that operational attributes of the media on which digital records are originally stored (such as speed of access) may constrain the intended functional behavior of such records, it may be necessary to preserve these attributes as well.

¹⁶⁰ The approach was first proposed in A. Michelson and J. Rothenberg, "Scholarly Communication and Information Technology: Exploring the Impact of Changes in the Research Process on Archives," *American Archivist*, Vol. 55, No. 2 (Spring), 1992 and discussed briefly in Rothenberg, 1995 (op cit, note 2). The most detailed discussion of it prior to this report appears in Rothenberg, 1999 (op cit, note 31).

D.1.4.8.1 Details of the emulation approach

The full, long-term implementation of the emulation approach would involve a number of steps: (1) Developing generalizable techniques for emulator specification that capture all of those aspects of the hardware computing environments to be emulated that are required to recreate the relevant behavior of current and future digital records, while facilitating the generation of emulators that will run on unknown future computers; (2) Developing techniques for saving the metadata needed to find, access, and recreate digital records in human readable form, allowing emulation techniques to be used for preservation; and (3) Developing techniques for encapsulating records, their attendant metadata, software, and emulator specifications in ways that ensure their cohesion and prevent their corruption.

The emulation approach involves encapsulating a number of items with each given digital record, consisting of three groups of information, as shown in Figure 8. There

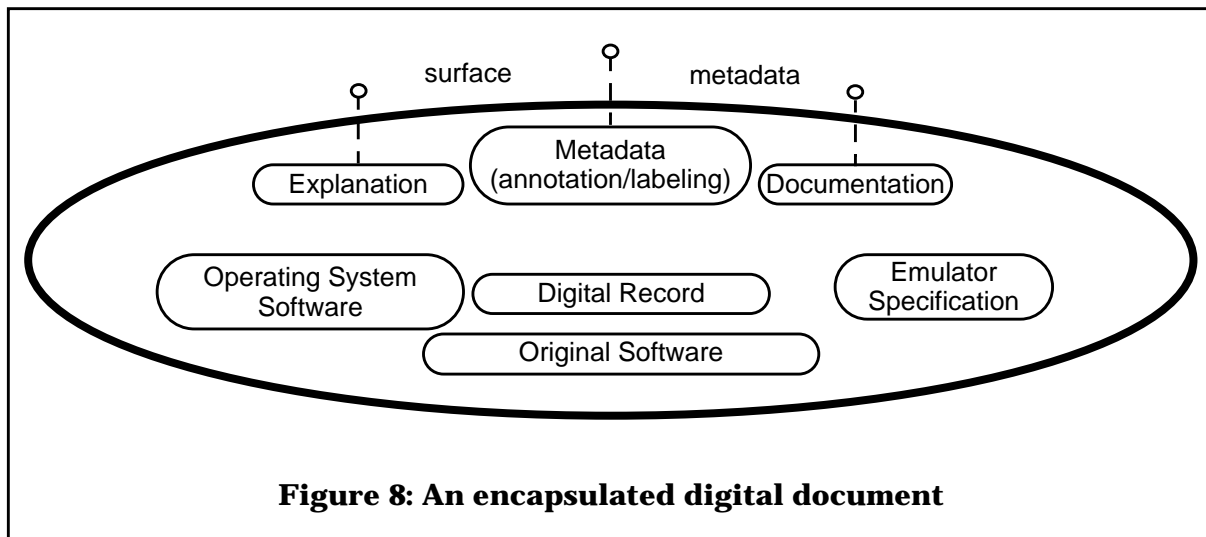


Figure 8: An encapsulated digital document

are a number of alternative ways of doing this, some of which would be safer (but more wasteful of storage) while others would involve somewhat more risk (but would use less storage). In practice, items that were required by many different records might be pointed to in centralized repositories rather than being replicated for each record.

Central to this encapsulation is the digital record itself, consisting of one or more files representing the original bit stream of the record as it was stored and accessed by its original software. In addition, the encapsulation contains the original software for the record—itsself stored as one or more files representing the original executable bit stream of the application program that created or displayed the record. A third set of files represents the bit streams of the operating system and any other software or data files comprising the software environment in which the record's original application software ran. These bit streams must be guaranteed to be copied verbatim when storage media are refreshed, to avoid corruption. This first group of

encapsulated items represents the original record in its complete software context: given a computing platform capable of emulating the record's original hardware environment, this information should recreate the behavior of the original record.

The second group of items in the encapsulation of a record consists of a specification of an emulator of the record's original hardware computing environment sufficient to allow the creation of such an emulator that will run on any conceivable computer (so long as it is capable of performing any computable function). This emulator specification cannot be an executable program, since it must be created without knowledge of the future computers on which it will run. Among other things, it must specify all attributes of the original hardware computing environment that are deemed relevant to recreating the behavior of the original record when running its original software under emulation. Note that only one emulator specification need be developed for any given hardware environment: a copy of it (or pointer to it) can then be encapsulated with every record whose software uses that environment. This second group of encapsulated items provides the key to running the software in the first group: assuming that the emulator specification is sufficient to produce a working emulator, the record can be read (i.e., accessed in its original form) by running its original software.

The final group of items in the encapsulation of a record consists of explanatory material, labeling information, annotations, metadata about the record and its history, and documentation for the software and (emulated) hardware included in the encapsulation. This material must first of all explain to someone in the future how to use the items in the encapsulation to read the encapsulated digital record. In order to serve this function, at least the top level of this explanatory material must remain human readable in the future, to serve as a "bootstrap" in the process of opening and using the encapsulation. This is one place where standards may find a niche in this approach: simple textual "annotation standards" (which might evolve over time) would provide one way of keeping explanatory material human readable. If translation of this explanatory material is required to keep it human readable (that is, if the annotation standards evolve), this might be performed when the encapsulation is copied to new media: we refer to this limited form of translation as "transliteration". Any such translation must be reversible without loss, to ensure (and allow verifying) that the explanatory material is not corrupted. (Note that these same techniques must be used to store emulator specifications, which must also remain human readable in the future.) Additional metadata in the encapsulation describes the original record, providing the equivalent of labeling information that should adhere to the record. Finally, additional metadata must provide historical context, provenance, management history, and administrative information to help manage the record through time.

Given a suitable emulator specification for the desired hardware computing environment (which need only be created once for all records whose software uses that environment), the process of preserving a digital record can be expressed by the sequence of steps "Annotate, Encapsulate, Transliterate, and Emulate". This means:

(1) create any annotations needed to provide context for the record and to explain how to open the encapsulation; (2) encapsulate all of the above required items with the record; (3) when necessary, translate annotations to keep them human readable; and (4) in the future, open the encapsulation, create the specified emulator (or access it if it has already been created and saved), and run it on a future computer to run the original software under emulation, thereby recreating the record.

D.1.4.8.2 Efficiency of the emulation approach

Though it may appear prohibitive to have to create and use an emulator to read each old record, four factors should be kept in mind. First, the inclusion of contextual annotation and metadata in the encapsulation makes it unnecessary to use emulation and run a record's original software in order to perform routine management functions on the record, such as copying it, filing it, distributing it, and (in some cases) finding and retrieving it: emulation is needed only when the record is to be "rendered" in order to be read in its original form or to be transcribed into some current "vernacular" (or "use-copy") form.¹⁶¹ Similarly, records may be translated into and stored in some such convenient "use-copy" form to allow them to be searched for and retrieved by content, without necessarily requiring that they be rendered in their original form via emulation.

Second, an emulator for a given obsolete hardware computing environment need be created at most once for each future type of platform on which it is required to run. Once created to run on a given new generation of computer, the emulator for a given obsolete hardware environment can be saved to be run on any such new computer whenever desired. Furthermore, if emulation of a given obsolete hardware environment is needed only rarely, it is not necessary to generate a new emulator for that environment to run on every new generation of computer. For example, if an emulator for obsolete hardware environment A runs in a subsequent hardware environment B that has itself become obsolete, but an emulator for environment B runs on a newer platform C, then the emulator for A can be run under the emulator for B on platform C¹⁶². Computer vendors have often provided successive layers of emulation of this kind in the past (as discussed in Section D.1.4.8.3 below), where it has been used effectively to avoid having to migrate programs and their files to every new generation of computer. Coupled with the fact that new, incompatible hardware computing environments are introduced fairly infrequently (compared, for example, with application software upgrades), this implies that emulators for new computers need be generated only relatively rarely.

¹⁶¹ See note 66.

¹⁶² If the need to emulate environment A subsequently becomes more common, its saved emulator specifications can always be used to generate an emulator to run directly on platform C (or any other platform).

Third, it should be possible to represent emulator specifications in such a way as to facilitate automating most of the process of generating emulators for new platforms from these specifications. For example, if emulator specifications are expressed in terms of an “abstract machine” that can be hosted on a future computer by implementing a minimal set of primitive instructions, then generating future emulators should be quite inexpensive. As computer science evolves, it should be possible to automate more and more of this process, until it ultimately requires little or no explicit human effort.¹⁶³

Finally—and most important of all—once an emulator specification is developed for a given hardware computing environment, all records, documents, and data that depend on *any* software that ran in that environment can be preserved simply by saving the bit streams of those records (along with those of their software and software environments). No further work is required to preserve any individual record: if and when a given record is actually accessed in the future, its original software is run under emulation, but even then no specific processing need be performed on that record. This represents potentially huge savings over schemes based on migration or standards, which require repeated translation of every individual record, *whether or not they are likely to be accessed in the future*. Therefore even if emulation is deemed unnecessary for certain classes of digital records, if it is done for any record, it can be used for all other records essentially for free. This gives the approach tremendous leverage.¹⁶⁴

D.1.4.8.3 Natural experiments related to emulation

A number of “natural experiments” have occurred that indicate the feasibility of using emulation to increase the longevity of programs and their documents. One prime example is the decades-old practice that hardware vendors have used to provide upward compatibility for their customers. Forcing users to rewrite all of their application software (and its attendant databases, documents, and other files) when switching to a new computer would make it hard for vendors to sell new machines. Many vendors (such as IBM) have therefore often supplied emulation modes for older machines in their new machines. The IBM 360, for example, included emulation modes for older IBM machines (such as the 7090) so that old programs written for those machines could still be run. In fact, some of those older machines themselves incorporated emulators for their own predecessors, in some cases back to the level of “plug-board” card processing computers, that were programmed by plugging wires

¹⁶³ Experiments such as those to be conducted using the testbed should reveal whether emulation is sufficiently promising as a preservation technique to warrant this kind of future formalization and development.

¹⁶⁴ Since this approach was first outlined in Michelson and Rothenberg, 1992 (op cit, note 160), it has received considerable attention, being cited as the only proposed approach that appears to offer a true solution to the problem of digital preservation A. Erlandsson, *Electronic Records Management: A Literature Review*, International Council on Archives’ (ICA) Study, 1996, <http://www.archives.ca/ica>, ISBN 0-9682361-2-X.

into patch boards. Programs that were several generations old were therefore routinely run on the 360 under several nested levels of emulation in this way. Not only was this an effective means of keeping old programs and their files usable, but they generally ran faster even under these successive layers of emulation than they had run on their original machines, since the 360 was so much faster than those older machines.

Apple Computer did something similar when switching from the Motorola 68000 processor series to the PowerPC by including an emulator for 68000 code; not only did this allow users to run all of their old programs on the new machine, but significant pieces of the Macintosh operating system itself were also run under emulation after the switch, to avoid having to rewrite them. Whether emulation is provided by a special mode using microcode or by a separate application program, such examples prove that emulation can be used to keep programs and their documents usable long after they would otherwise have become obsolete.

Another example of the use of emulation for preservation is in the “retro-computing” community, whose members create emulators for obsolete video game platforms and other old computers. There are numerous World Wide Web sites listing hundreds of free emulators of this kind that have been written to allow old programs to be run on modern computers. A particularly interesting example of this phenomenon is the MAME (Multiple Arcade Machine Emulator) system, which supports emulation of a large number of different platforms, demonstrating that emulation can be cost-effective for a wide range of uses.

All of these examples consist of emulators that run on existing hardware platforms, so they do not address the problem of specifying an emulator to run on a future, unknown computer; yet they prove that emulation is an effective way of running otherwise obsolete software.

A somewhat different example of the use of emulation is in designing new computing platforms. Emulation has long been used as a way of refining new hardware designs, testing and evaluating them, and even beginning to develop software for them before they have been built. Although this is not an example of using emulation to preserve old documents, emulators of this kind might be a first step toward producing the emulator specifications needed for the approach proposed here. Hardware vendors might be induced to turn their hardware-design emulators into products that could satisfy the emulator scheme’s need for emulator specifications. Hardware selection decisions by government agencies and other recordkeeping organizations might even be based on the availability of such emulator specifications along with the hardware.

D.1.4.8.4 Unanswered questions about emulation

In addition to the research required to develop the emulation approach (discussed in Section D.1.4.8.1 above), there remain a number of unanswered questions about emulation.

The most obvious of these is how well the approach can reproduce the many aspects of an obsolete hardware computing environment that might be required by a range of authenticity criteria. Digital documents are in general designed to be used on a range of hardware platforms, for example, having processors of different types, versions and speeds, displays of different sizes, shapes, resolution and color characteristics, storage devices of different speeds and capacities, and input devices of various kinds, such as keyboards, mice, trackballs, trackpads, joysticks, etc. It is therefore presumably not necessary (at least in most cases) to duplicate the exact characteristics of any specific combination of such devices in order to recreate an authentic digital record, since its original use involved a variety of such environments. Nevertheless, it is important to understand just what aspects of hardware computing environments emulator specifications would need to capture.¹⁶⁵

In addition, there is the issue of how to endow emulation environments with the ability to extract use-copies or other transcribed versions of documents from their emulated forms. For example, if the original software for a digital record displayed a text document as an image without any encoding of the underlying text, can a textual use-copy be extracted from this in the future, when the image is displayed under emulation? In this case, the emulated image might be scanned by software running outside the emulation on the future computing platform, but for this to be possible, the emulation environment must be capable of presenting the documents it renders to the computing environment in which it is running. In general, the emulation environment must provide “hooks” to allow such extraction, based on a flexible model of the kinds of content that it might be desired to extract.¹⁶⁶

As discussed above, emulation requires encapsulation of bit streams and related explanatory documentation in order to enable access to digital records in the future. Because emulation requires running original application and operating system software in the future, the bit streams of these programs in particular must be copied faithfully for the scheme to work. Any corruption of these programs or their data files may result in fatal bugs when they are run under emulation.¹⁶⁷

¹⁶⁵ Beyond these device characteristics, there is the question of whether and when it may be necessary to emulate specialized co-processors, such as floating point, memory-management, graphics, or sound processing chips or cards, as well as unusual input/output devices such as datagloves, data-capture instruments, head-mounted displays, etc.

¹⁶⁶ Fortunately, the relevant types of content depend on the record itself (and its original use) rather than some unknown future; it should therefore be possible to define the emulator specification for the record’s original hardware environment to include appropriate “hooks” for future extraction limited to those kinds of content that make sense for the record in its original context.

The pragmatic and legal issues surrounding the acquisition and use of proprietary software and hardware descriptions pose a different kind of challenge to the emulation approach. As noted above (in Section D.1.2.3), there are many unresolved issues surrounding copyright and intellectual property in the digital domain. In particular, it may be difficult to obtain the rights to all of the information required for emulation. This may ultimately have to be solved in the legislative or policy arena, though it is worth noting that—in this case at least—time appears to be on the side of the archivist.¹⁶⁸

D.2 Comparing alternative preservation approaches

Despite their shortcomings, the alternative preservation approaches discussed above are all potential candidates for experimentation in the testbed proposed in Annex C. As a way of beginning to compare these alternatives, we offer the following informal “cost-effectiveness” analysis. Both the factors and the arithmetic used to combine them here are highly notional, as are the actual values supplied. The formulation is intended to be merely suggestive:

$$\begin{aligned} \text{Cost} = & \text{InitialCost.1} + (\text{RecurringCost.1} * \text{FrequencyOfRecurrence.1}) \\ & + \text{InitialCost.2} + (\text{RecurringCost.2} * \text{FrequencyOfRecurrence.2}) \\ & \vdots \\ & + \text{InitialCost.n} + (\text{RecurringCost.n} * \text{FrequencyOfRecurrence.n}) \end{aligned}$$

$$\text{Cost.i} = f(\text{effort.i, time.i, money.i, OtherResources.i})$$

$$\text{Effectiveness} = (\text{FractionOfRecordsRetained} * \text{ValidationMeasure}) ^ \text{AvFreq}$$

(where AvFreq is Average frequency of recurrence)

$$\text{FigureOfMerit} = \text{Effectiveness} / \text{Cost}$$

¹⁶⁷ The emulation specifications themselves must also be copied faithfully, but since these are expected to be textual descriptions or high-level programs accompanied by documentation, they may be able to tolerate a certain degree of corruption, since they may incorporate sufficient redundancy to allow repairing minor damage. Such redundancy (i.e., error-correction techniques) can and should be designed into these emulator specifications. However, the original software associated with a digital record will consist of executable, binary programs that were created by vendors with no particular motivation to produce redundant, self-correcting code. Errors introduced into these bit streams may therefore pose a more serious threat to the emulation approach. It should be possible, however, to add redundancy to such bit streams (as well as to the digital records themselves) when saving them in a preservation system: this can be done using replication, digital signature (“checksum”) techniques, error-correction encoding, etc.

¹⁶⁸ That is, the need to use obsolete software and descriptions of obsolete hardware environments arises directly from their obsolescence, which implies that they will no longer be of commercial value by the time they are needed to recreate obsolete records. By definition, if software or hardware is still viable, it is not necessary to use emulation to retrieve saved digital records, whereas when the software or hardware becomes obsolete, it should no longer have any market value. Though market value may not be the only criterion for determining intellectual property rights, it may be possible to reach some sort of legislative compromise between these factors.

We represent the cost of applying a preservation approach to a set of digital records as the sum of a set of initial costs and a set of terms representing recurring costs, each weighted by its own frequency of recurrence. Since some approaches will have to treat different types of digital records differently, each initial cost represents the cost of applying the approach to one type of digital record when that type first enters the preservation process. Similarly, the multiple terms for recurring costs represent the fact that each type of record may have its own independent recurring cost, recurring with its own independent frequency of recurrence. The total number of such independent costs (i.e., the number of different types of records that have to be treated differently) is n . The frequency of recurrence of any of these costs will not necessarily be constant but is meant to represent an estimated average over some reasonably long period, such as 50 years.

Each type of cost is considered to represent some combination of the effort, time, money, and other resources required to perform some aspect of the required preservation process. As a first approximation, these different elements of cost can be expressed in monetary terms and combined by adding them together.

Effectiveness is intended to express how well the preservation process retains digital records. Its first factor represents the fact that some preservation approaches may not be able to preserve all records; the second factor represents the fact that even those records that are preserved will in general be preserved only to some extent, as represented by a hypothetical validation measure. Each of these factors will be less than 1, as will their product. The loss represented by this product will be compounded each time the process is applied, which is crudely represented by raising the loss per application to the power of the average frequency with which the process is applied. This effectiveness measure can be computed separately for each type of record that is treated differently by the preservation process (that is, for each index, i), in which case $\text{FrequencyOfRecurrence}_i$ from the cost formula should be used in place of AvFreq .

The final figure of merit derived from the above consists of Effectiveness divided by Cost. To compute this, we assign speculative input values from the symbolic value set

{low, medium, high}

and we allow the result to take on the values

{low, LowMedium, medium, HighMedium, high}

performing informal symbolic arithmetic on these symbolic values. Performing these symbolic computations for three sample preservation approaches produce the following results.

For a preservation approach based on standards, we estimate:

InitialCost = high

RecurringCost.1 = medium

Frequency.1 = low

RecurringCost.n = medium

Frequency.n = low

n = medium

Fraction of records retained = low

Validation measure = medium

Average frequency = low

FigureOfMerit = $(\text{low} * \text{medium})^{\text{low}} / (\text{high} + \text{medium} * (\text{medium} * \text{low}))$
= HighMedium / LowMedium = medium

Whereas, for a preservation approach based on migration, we estimate:

InitialCost = medium

RecurringCost.1 = high

Frequency.1 = medium

RecurringCost.n = high

Frequency.n = medium

n = high

Fraction of records retained = medium

Validation measure = medium

Average frequency = medium

FigureOfMerit = $(\text{medium} * \text{medium})^{\text{medium}} / (\text{medium} + \text{high} * (\text{high} * \text{medium}))$
= LowMedium / HighMedium = low

And finally, for a preservation approach based on emulation, we estimate:

InitialCost = low

RecurringCost.1 = low

Frequency.1 = low

RecurringCost.n = low

Frequency.n = low

n = low

Fraction of records retained = high

Validation measure = high

Average frequency = low

FigureOfMerit = $(\text{high} * \text{high})^{\text{low}} / (\text{low} + \text{low} * (\text{low} * \text{low}))$
= high / low = high

While these results should not be taken too seriously, they serve as an expression of our initial hypotheses about the relative values of alternative preservation approaches. Experimentation with the testbed should provide more empirical results.

D.3 Scope and limitations of the study

This study was conducted in the rich context of archival recordkeeping and preservation. Yet its limited resources made it necessary to identify and focus on the most relevant aspects of this context. It quickly became apparent that the scope of the study would have to be carefully defined to fit within its short timeframe and budgetary constraints. Much of the first phase of the study was therefore spent trying to understand the scope issues and determine a useful and appropriate focus for the study. This scoping effort led to the following restrictions.

D.3.1 Focus on technical aspects of digital archival preservation

The primary scope restriction was that the study should focus on the technical aspects of digital archival preservation, rather than on the intellectual or organizational issues surrounding preservation or on ancillary recordkeeping issues such as those concerning the maintenance of contextual metadata or other archival concerns such as selection, description, or access.

Nevertheless, many of these other issues have a direct bearing on digital preservation, making it impossible to eliminate them from consideration entirely. For example, it is impossible to devise methods for preserving digital archival records without understanding what such records are, which of their attributes are relevant to preserving them as authentic, meaningful records, what organizational and administrative impediments may exist to specific technical approaches, or what requirements archival selection and description may place on the ways that digital records can be represented. Therefore, although such issues were not strictly within the scope of this study, they are discussed where appropriate throughout this report.

It was also considered important to take into account emerging trends in electronic commerce (“e-commerce”) and electronic (or “digital”) government in The Netherlands and elsewhere, as well as the evolving vision of a national “Corporate Archival Database” for the country that would encompass cultural heritage projects, the Royal Library, and holdings of various museums throughout The Netherlands. However, scope constraints prevented the study from performing more than a cursory analysis of these efforts, leading to the tentative conclusion that while they are potentially relevant to the digital preservation enterprise, they are not yet well enough defined to provide specific requirements, constraints, or opportunities for that endeavor, though they may offer the potential for synergistic interaction in the future.

Even within the realm of technical preservation, many issues have had to be excluded from the scope of the study and are mentioned only in passing. In particular, technical

issues surrounding the “ingestion” of records into a digital preservation system and issues concerned with searching for digital records, creating and using “finding aids” (metadata used in searching), and retrieving records are not discussed in any detail. Access of records is identified as an essential aspect of preservation, but it too is discussed only superficially. These and other technical aspects of preservation (such as deaccessioning or disposing of records, “redacting” or excising parts of records that are unreleasable for security or other reasons, and handling the large volume of digital records that can be expected in the future) are important and must be analyzed in depth as part of the process of designing a fully operational preservation system, but they have had to be excluded from the central focus of this study.

D.3.2 Focus on mainstream problems that generalize

In order to make the study relevant to the real problems of the National Archives and the Dutch Government, its scope was limited to aspects of the problem that are either already occurring or are expected to occur in the near future with sufficient frequency to have significant impact. The rapid evolution of information technology makes this problematic, since (to paraphrase the safety warning that appears on wide-angle automotive mirrors) problems in the future are closer than they appear. Hypertext or network-based documents, object-oriented databases, or records embodying voice annotations, video clips, computer animation, or active, executing software may seem to be in the far future, yet they may come into use sooner than most people expect. Still, in order to restrict the scope of the study, such “futuristic” forms of digital records were largely excluded from consideration.

Nevertheless, the scope was extended to include enough challenging types of digital records to ensure that the results of the study would have a good chance of generalizing to the more exotic types excluded above. In particular, with the resounding endorsement of the study’s Advisory Group, it was decided to include non-textual forms of records, such as graphics and databases, in an attempt to achieve generalizability. Without this extension (for example, if its scope had been restricted to textual digital records) it was felt that the study’s results would not apply to a sufficiently large number of the problems that are already being encountered.