The InterPARES Preservation Model: A Framework for the Long-Term Preservation of Authentic Electronic Records

William E. Underwood

Abstract.

This paper describes the InterPARES Preservation Task Force's analysis of the problem of preserving electronic records.[1] The InterPARES Preservation Model provides a generic preservation strategy (or framework) for preserving authentic electronic records. Within that framework, a variety of preservation strategies can be developed by archival institutions that are dependent on the characteristics of the selected, transferred and accessioned records, institutional requirements, and the current and changing state of information technology. To refine and validate the Preservation Model, walkthroughs of the model are being conducted using information from case studies. Results of a walkthrough are described. It is demonstrated that the Preservation Model provides a framework for implementing procedures that satisfy the Authenticity Task Force's (ATF's) Baseline Requirements for Supporting the Production of Authentic Copies of Electronic Records. The model also includes an activity for using the ATF's Benchmark Requirements to assess the presumption of authenticity that can be accorded a creator's records. An example is given from the case study of the kinds of information that would be required of the creator's records to determine whether they could be presumed authentic.

The Problem of Preserving Electronic Records

The rapid obsolescence of computing technologies creates difficulties for those concerned with the long-term preservation of records in digital form. The potential need to migrate these records across hardware and software technologies raises questions related to the records' authenticity. How can one ensure that sets of digital records have not been intentionally or inadvertently modified? How can one ensure that long-term preservation methods do not compromise the authenticity of digital records?

The Preservation Task Force's research objective was to develop a generic solution to the problem of preserving authentic electronic records. The IDEF0 modeling notation

---

[1] This paper was completed as part of the InterPARES Research Project (http://www.interpares.org). This project, which is investigating the preservation of permanent authentic records in electronic record-keeping systems, brings together an interdisciplinary research team drawn from National Archives and universities in North America, Europe, and Asia. The project has received substantial support from the Canadian Social Science and Humanities Research Council Major Collaborative Research Initiative (MCRI), the National Historical Publications and Research Commission in the United States, and the Italian National Research Council. The author of this paper is a member of the American InterPARES Research Team and the InterPARES Preservation Task Force. This research was also supported in part by the Electronic Records Archives Program of the National Archives and Records Administration.

and methodology was used to represent the problem and the results of our analysis of the problem.[2] At the most abstract level, the IDEF0 context diagram in Fig. 1 represents the problem of preserving authentic electronic records.

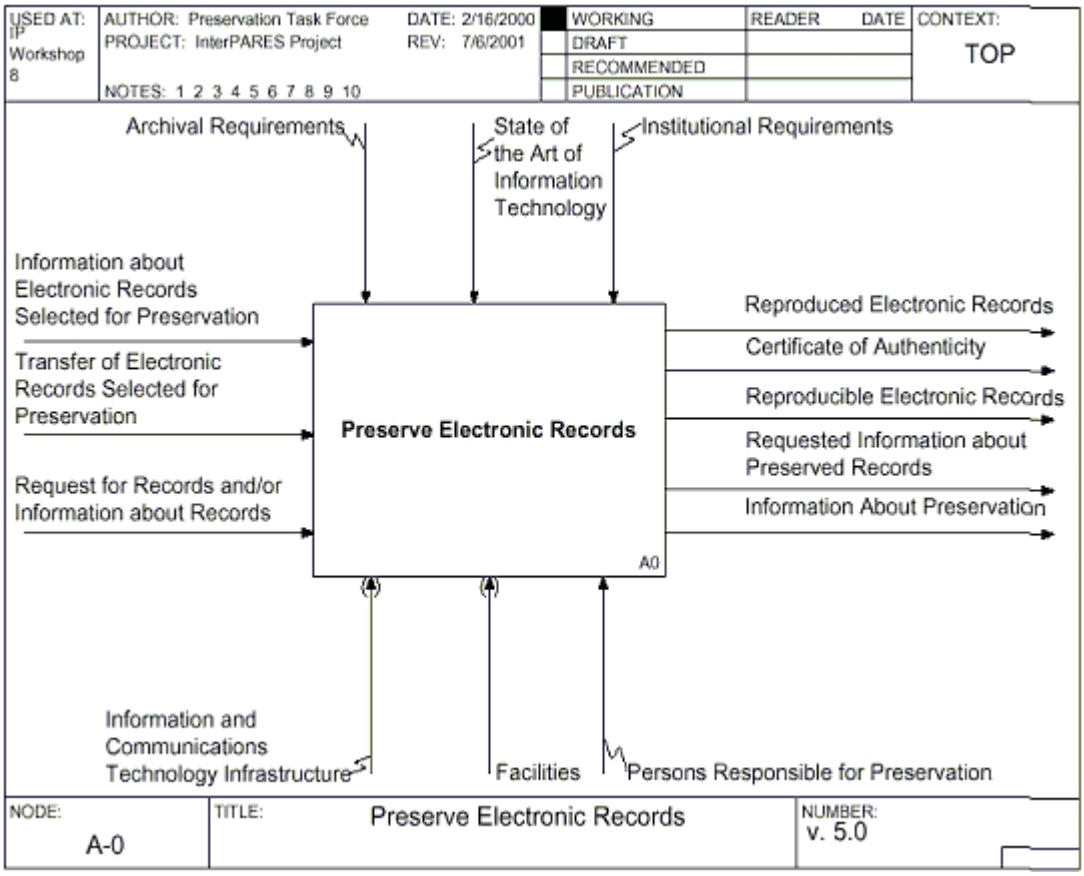

Figure 1. IDEF0 Representation of the Preservation Problem.

Given *Information about Electronic Records Selected for Preservation* and *Transfers of Electronic Records*, the goal is to preserve these electronic records so that given a *Request for Records* or a *Request for Information about Records*, the requested records can be reproduced, and information about those records and preservation actions on those records can be provided. The box in the center of this diagram represents the general activity of *Preserving Electronic Records*. The labeled arrows entering the box from the left represent the inputs to the activity. The activity transforms the inputs to the outputs, which are shown as labeled arrows leaving the right side of the activity box. The labeled arrows entering the top of the activity box represent controls that regulate the activity, for example, *Institutional Requirements* govern the preservation of electronic records. The physical resources required to perform this activity are represented as labeled arrows entering the bottom of the activity box. They include *Information and Communication Technology*, *Facilities* and *Persons Responsible for Preservation*. The problem is analyzed from the point of view of

[2] US Department of Commerce. FIPS Pub 183, *Integration Definition for Function Modeling* (*IDEF0*), 1993.

persons responsible for preservation, not those archivists responsible for appraisal, review, description or access.

The preservation problem was analyzed and decomposed into the four subproblems shown in Fig. 2.

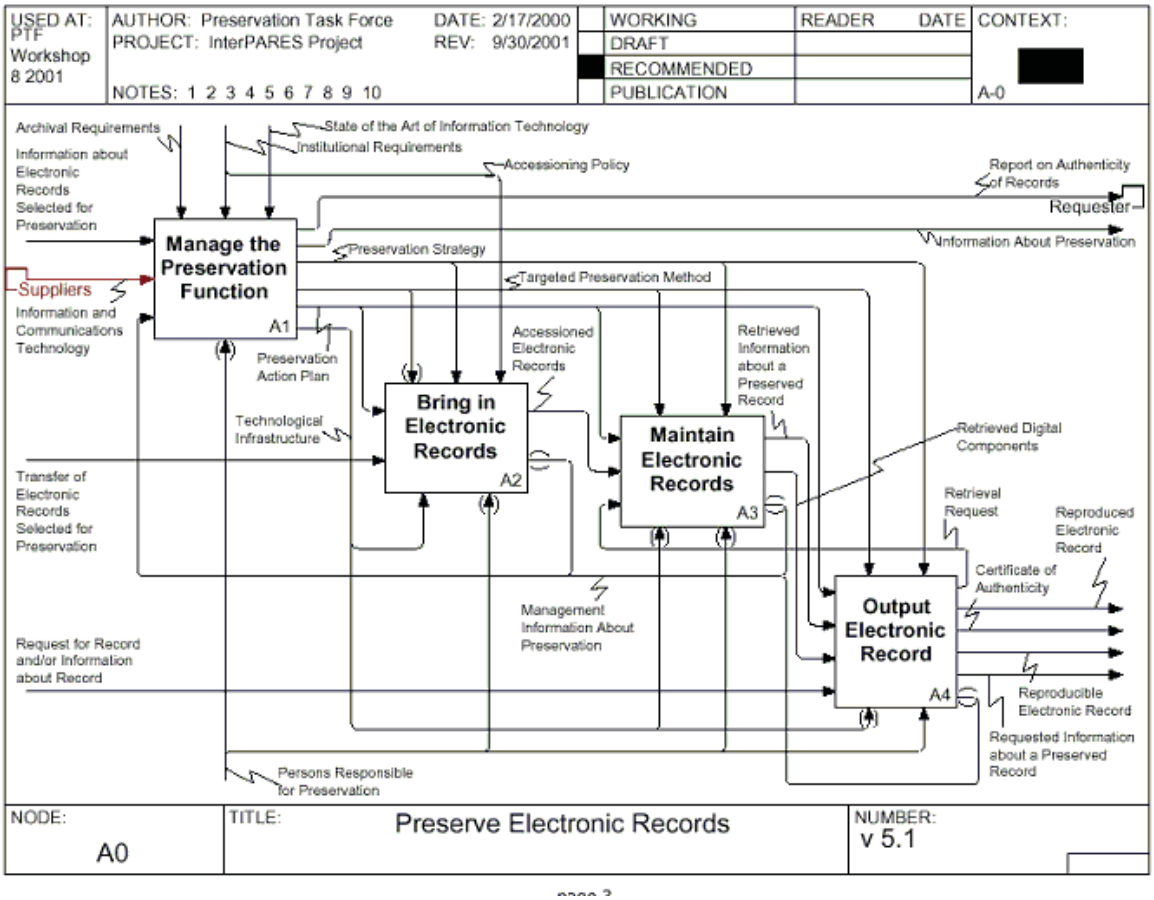

Figure 2. Decomposition of the Preservation Problem.

When there is a *Request for a Record and/or Information about a Record*, subproblem A4 is to achieve the goals of *Reproduced Electronic Records* and *Requested Information about a Preserved Record*, given *Retrieved Information about a Preserved Record*, *Retrieved Digital Components*, and *Targeted Preservation Methods* (*for Reproducing the Records*). The subgoals of having *Retrieved Digital Components* and *Retrieved Information about a Preserved Record* can be achieved, if we have maintained (activity A3) the *Accessioned Electronic Records*, and *Planned Action Plans* have been executed using *Targeted Preservation Methods*. The subgoal of having *Accessioned Electronic Records* can be achieved, if we can bring in (activity A2) the *Transferred Electronic Records Selected for Preservation* and *Preservation Action Plans* can be executed using *Targeted Preservation Methods* to bring the *Transferred Electronic Records* into compliance with the preservation strategy. The subgoals of having *Preservation Action Plans* and *Targeted Preservation Methods* can be achieved, if we have *Information about the Electronic Records Selected for Preservation* and

preservation decisions are made based on *Archival Requirements*, the *State of the Art of Information Technology*, and *Institutional Requirements*. Each of the four subproblems shown in this diagram was analyzed and decomposed into subsubproblems. The decomposition of problem A1, *Manage the Preservation Function*, and problem A3, *Maintain Electronic Records*, will be described.

Subproblem A1, *Manage the Preservation Problem*, was analyzed and decomposed into the four subproblems shown in Fig. 3.

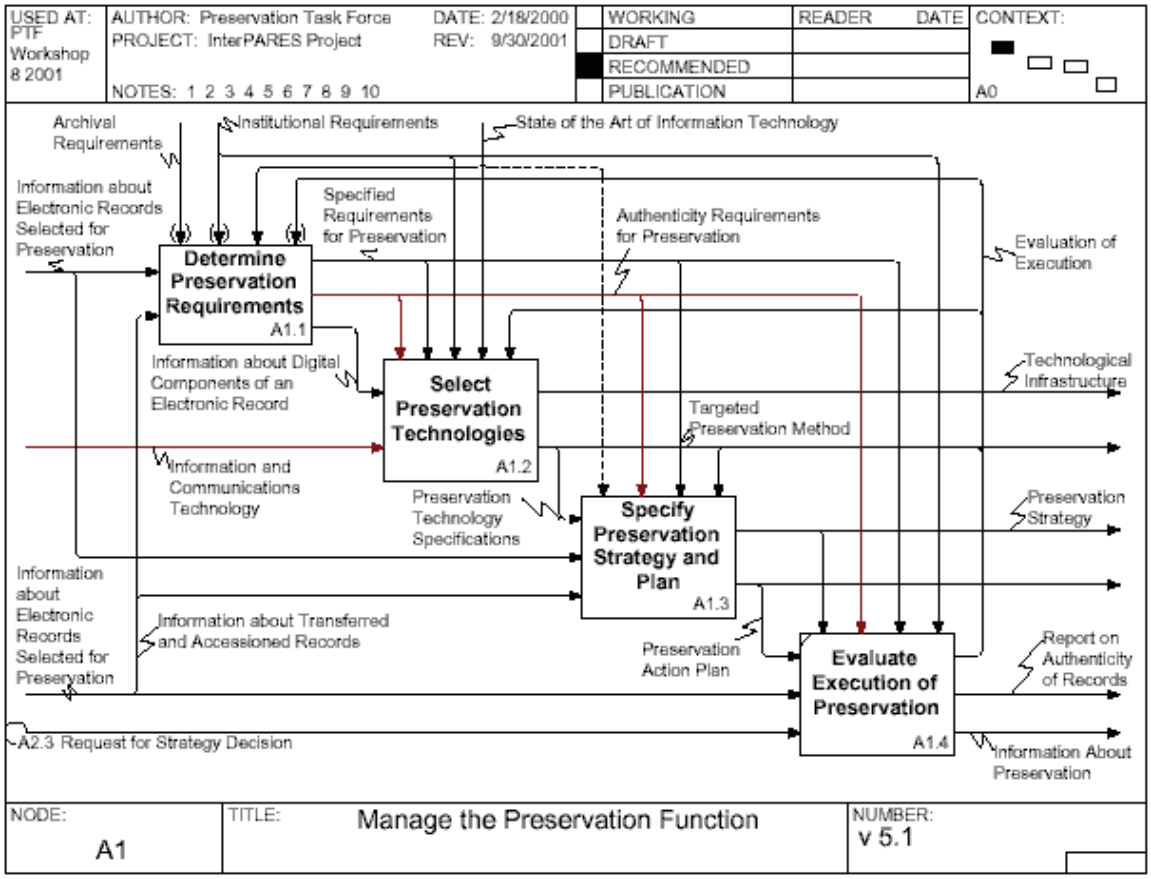

Figure 3. Decomposition of the Problem Manage the Preservation Function.

Helen Forde quoted Trudy Peterson as stating that preservation should be viewed "as a program [*process*] to be managed, not a problem to be solved."[3] The InterPARES Preservation Model reflects this keen observation. The preservation model is an activity or process model. In activity A1, *Manage the Preservation Function*, preservation choices are made and strategies articulated. It is also in this activity that feedback from the preservation process is assessed and preservation strategies, plans and methods are refined.

[3] Helen Forde, Preservation of traditional materials: paper, parchment, bindings and seals**.** This proceedings.

The problem of *Managing the Preservation Function* is solved by solving the subproblems of *Determining Preservation Requirements*, *Selecting Preservation Technologies*, *Specifying Preservation Strategy and Plan*, and *Evaluating the Execution of Preservation*. Each of the problems was analyzed and decomposed into subproblems. For instance activity A1.3, *Specify Preservation Strategy and Plan* was decomposed into the three subproblems shown in Fig 4.

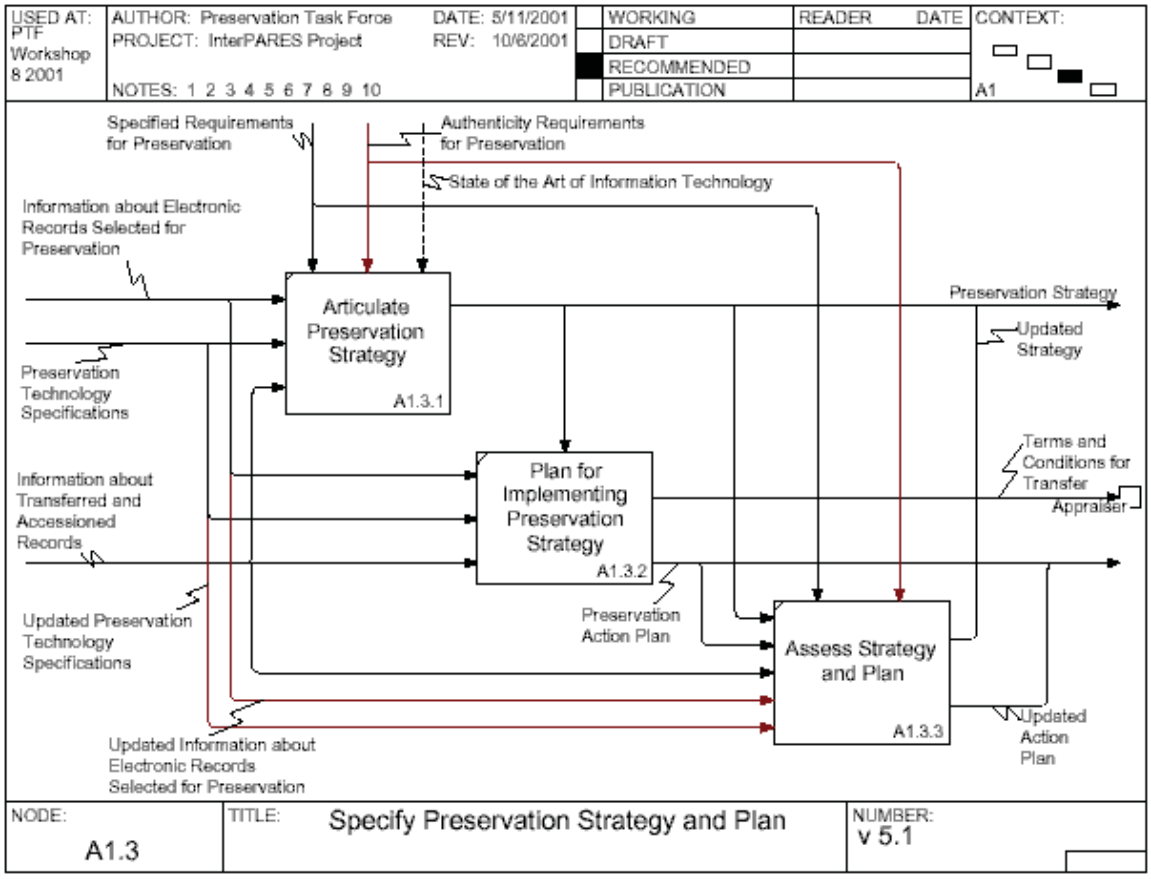

Figure 4. Decomposition of the Problem, Specify Preservation Strategy.

The solution to the first problem, *Articulate Preservation Strategy*, involves the choice, by the person responsible for preservation, of a preservation strategy for overcoming the problem of technological obsolescence of the hardware and software used to create the records selected for preservation. The current State of the Art of Information Technology might currently indicate that the possible preservation strategies for overcoming obsolescence of the computer platform include:

1. Migrate software viewers for digital components to current computers, that is to say, to purchase a new viewer for the platform or to recompile or reprogram the viewer source code for the new platform.
2. Migrate (or convert) an obsolete format to a current format.
3. Convert digital components in obsolete proprietary formats to standard formats, for example, to convert a dBase IV file to SQL.

4. Convert digital components in proprietary formats to descriptions in standard markup and presentation languages such as XML, XML Schema, and XSL-FO.
5. Emulate the obsolete computer processors, storage and display devices on current processors, storage and display devices, so that the original software can be used to reproduce records.

The solution to the second problem, *Plan for Implementing Preservation Strategy*, would be constrained by the Preservation Strategy chosen for this body of records and would produce *Terms and Conditions for Transfer* and *Preservation Action Plans*. The third problem is to *Assess the Preservation Strategy and Plan* for a specific body of records and possibly to update this strategy and plan.

Returning to the high-level problem decomposition, problem A3, *Maintain Electronic Records*, was decomposed into the three subproblems shown in Fig. 5.

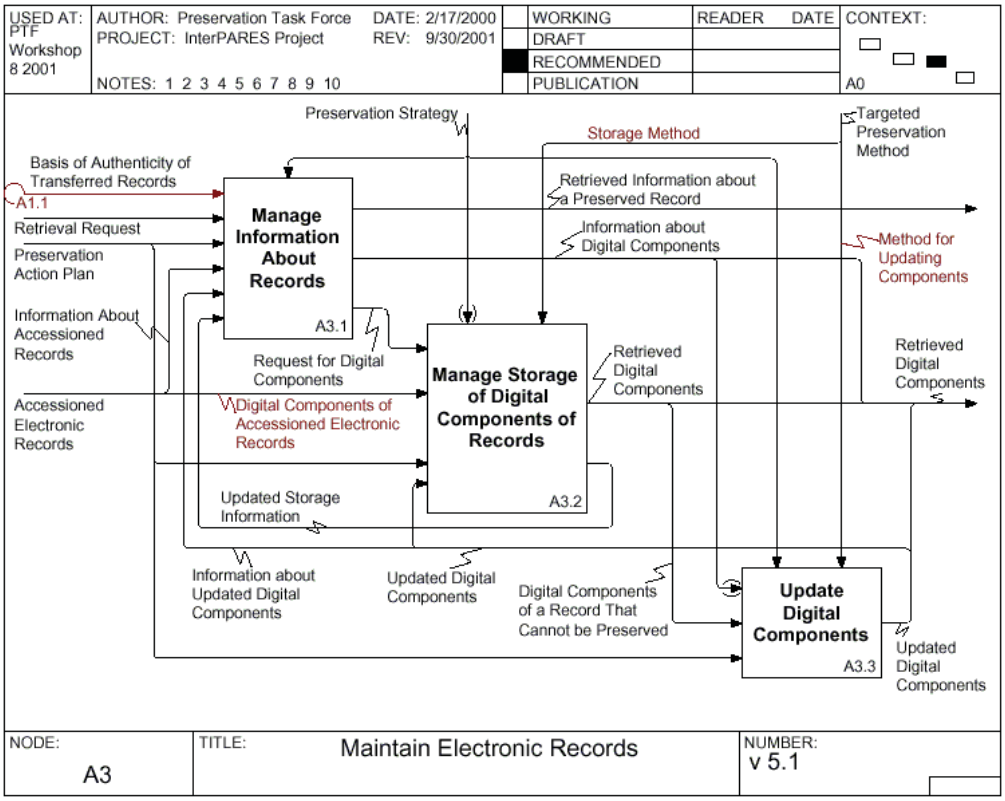

Figure 5. Decomposition of the Problem of Maintaining Electronic Records.

Problem A3.1, *Maintain Information about Electronic Records*, can be solved through the use of a database management system that supports storage, update and retrieval of information about accessioned electronic records. Problem A3.2, *Manage Storage of Digital Components of Records*, can be solved with an archival storage system that supports storage and retrieval of the digital components of accessioned electronic records. Problem A3.3, *Update Digital Components*, has as its goal that records be reproducible from their digital components. However, the obsolescence of the file formats of the digital components due to new computer hardware, system

software or application software places the records at risk of not being reproducible. Problem A3.3, *Update Digital Components*, was analyzed and decomposed into three alternative subproblems shown in Fig. 6.

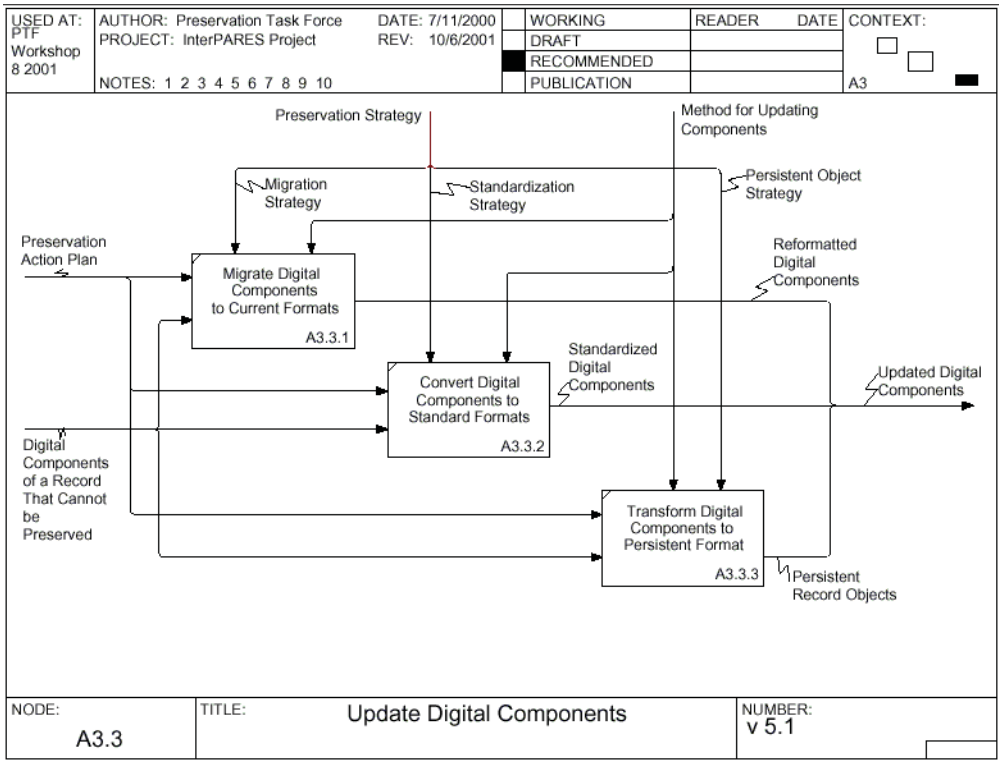

Figure 6. Decomposition of the Problem of Updating Digital Components.

Activity A3.3.1 solves problem A3.3, the problem of *Updating Digital Components* to overcome technological obsolescence, by using a preservation method that migrates digital components in obsolete formats to current file formats. Activity A3.3.2 solves the problem of *Updating Digital Components* by using a preservation method that converts digital components represented in proprietary or obsolete file formats to standard file formats. Activity A3.3.3 solves problem A3.3 by using a preservation method that transforms digital components in a proprietary, obsolete, or standard format into descriptions of the record's documentary and physical form in a standard markup language, such as the Extensible Markup Language (XML) and the Extensible Stylesheet Language for Formatted Objects (XSL-FO).[4]

The process of decomposition is continued until all subproblems have a solution in terms of actions that can be performed by a person, by computer programs, or by a combination thereof. This decomposition can be represented as a tree with the preservation problem (or context diagram) at the root of the tree and with the leaves at the ends of the branches of the tree representing solutions to the subproblems.

[4] For a description of persistent objects and persistent archives see A. Rajasekar, R. Marciano, and R. Moore, Collection-based persistent archives, San Diego Supercomputer Center, www.sdsc.edu/NARA/

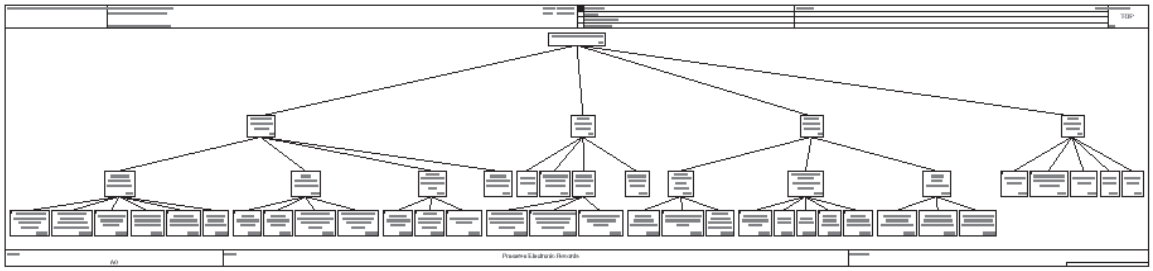

Figure 7. Problem Reduction Tree for the Preservation Problem.

The InterPARES Preservation Model is a conceptual specification of the problem of preserving authentic electronic records and a framework for solving it. It is a reference model.

Walkthrough of the Model using Case Study Data

The InterPARES preservation model is a generic model of the process of preserving authentic electronic records. If the model included specific preservation decisions, the generality of the model would be compromised. On the other hand, it is intended that it provide a framework for making and carrying out preservation decisions. How can archivists know that it is an effective framework for guiding management decisions and implementing preservation strategies?

Walkthroughs using case data are an effective way to test whether a model, design, program code, or user interface achieve what is intended and to improve the quality of the product.[5] A walkthrough is a peer group review of any information system product. A walkthrough of an activity model, such as the preservation model, is concerned with the functionality of the system. Walkthroughs can also be used to determine whether an activity model or design meets functional or nonfunctional requirements, such as the Baseline Requirements Supporting the Reproduction of Authentic Electronic Records. To demonstrate that the Preservation Model applies to specific cases of electronic records selected for preservation, to refine and validate the Preservation Model, and to demonstrate that the preservation model satisfies the Baseline Requirements a series of walkthroughs is being conducted.

The walkthrough team consists of a *presenter*, who "puts on the table" the model being reviewed; *reviewers*, who have a good understanding of the model, ask questions of the case study expert to identify data corresponding to inputs and outputs of the activities, and raise issues and suggested solutions to problems; a *case study expert*, who answers questions posed by the reviewer about data from the case study; and a *secretary*, who records the discussed facts and issues and distributes the minutes.

---

[5] E. Yourdon. *Structured Walkthroughs*, 4th Ed., Englewood Cliffs, NJ: Yourdon Press, 1989. E. Freedman and G. Weinberg. *Handbook of Walkthroughs, Inspections and Technical Reviews*, 3rd Ed., New York: Dorsett Home Publishing, 1990.

The method used in the walkthrough is to iteratively step through each of the lowest-level activities in the Preservation Model:

(1) Reviewing the activity definition and the input, output and control definitions.
(2) Identifying data elements of labels on input and output arrows.
(3) Defining the transformation of inputs to outputs.
(4) Determining values of the data elements that are related to the specific body of records in the case study.
(5) Recording the results and any problems or issues that arise and suggesting possible solutions.

The case study used in the initial walkthrough was InterPARES Case Study 26, The New York Workers' Compensation Board (NYWCB) Electronic Case Folder System (ECFS).[6] One of the results of the walkthrough was to identify the data elements of objects created by the activities of the preservation model.[7]

The *Terms and Conditions for Transfer* are the specifications governing the transfer to the preserver of a body of electronic records selected for preservation. Fig. 8 shows the kinds of information that occur in the Terms and Conditions for Transfer with sample data from the case study.

Record creator's name: New York State Workers' Compensation Board
Transfer agent's name: John Doe, Records Manager
Identification of records:
Title: Electronic Case Folder System
Description: Series of case files for adjudicating benefits of disabled workers.
Document Types: Claims for Benefits, Employer's reports of accidents and illness, Correspondence, Medical Reports, Insurance Carrier's Reports
File Format: Multi-page TIFF
Volume: 300,000 open cases
Data structure: Relational Database Schema
Scheduled Transfer Date: To be determined
Medium or channel of transfer: DLT Tape
Technical Conditions for Transfer:
MD5 hash code of all transferred files for integrity check,
All documents converted to TIFF Multi-page format,
Metadata schema represented in SQL
Information needed to support a presumption of authenticity:

Figure 8. Elements of the Terms and Conditions for Transfer.

The last item in the Terms and Conditions for Transfer, Information needed to support a presumption of authenticity, refers to the Authenticity Task Force's set of Requirements for Assessing a Presumption of Authenticity of the Creator's Records.[8] Fig. 9 shows in the left column, the name of the requirement, and in the right column,

[6] Preservation Task Force, A Walkthrough of the PTF IDEF0 Model for Preserving Electronic Records. Appendix to the InterPARES Final Report.
[7] The walkthrough was conducted using Version 5.1 of the Preservation Model. Revisions were suggested that were incorporated into Version 6.0 which is included as an appendix to the InterPARES Final Report.
[8] Heather McNeil, InterPARES 1 Project, This proceedings.

examples of the kinds of information from the case study that would be needed for the preserver to assess the degree to which the creator's electronic records could be presumed authentic.

Benchmark Authenticity Requirement	**Information identified at the time of appraisal that is needed to support a presumption of authenticity**
A.1.a Identity of the record A.1.a.i Name of author Name of addressee A.1.a.ii Name of action or matter A.1.a.iii Chronological date A.1.a.iv Expression of Archival Bond A.1.a.v Indication of attachments	 The ECFS data model permits the association of author's name, addressee, name of action or matter, and chronological data, but does not actually capture it. When documents are imported by FileNet, a case file is ordered by document number. Document preparation and mail transmittal preparation rules address how attachments are kept in the case folder.
A.1.b Integrity of the record A.1.b.i Name of Handling Office A.1.b.ii Name of OPR A.1.b.iii Indication of types of annotations A.1.b.iv Indication of technical modifications	 NYWCB NYWCB FileNet supports annotations, but they are not used. Paper documents were scanned into document images in TIFF 6 format and maintained on WORM disks.
A.2 Access Privileges	Access to the ECFS is controlled via passwords, job titles, workgroups, geographic location and business need.
A.3 Protective Privileges: Loss and Corruption of Records	There are backup copies of the WORM disks and transaction logs.
A.4 Protective Privileges: Media and Technology	WORM Disks are guaranteed for over 100 years.
A.5 Establishment of Documentary Forms	Each form is described in a procedural manual that is managed in Lotus Notes.
A.6 Authentication of Records	Authentication of document images in a case file is occasionally required in the adjudication process. The documents images are presumed authentic because they are scanned images of paper documents and they are used in the normal course of business.
A.7 Identification of Authoritative Record	The document images are the authoritative record unless the paper file is still available.
A.8 Removal and Transfer of Relevant Documentation	There has not yet been a transition of active records to inactive status, which would involve the removal of records from the electronic system.

Figure 9. Information Needed to Assess a Presumption of Authenticity of the Creator's Records.

Requirement A1 prescribes that the identity of the electronic records be recorded in terms of name of author, name of addressee, name of action or matter and the chronological date of the record. While the Electronic Case Folder System permits the

association of values of these attributes with a document image in the case folder, the ECFS does not currently capture each of these values. Consequently, the preserver's degree of belief that the first requirement was met would be very low. The guidance to the Record Creator would be that for the preserver to presume that the document images in the case folder were authentic, the ECFS should capture in the metadata the name of author, name of addressee, name of action or matter and the chronological date of the document image.

In the walkthrough of activity A1.3.1, *Articulate Preservation Strategy*, the first preservation strategy was chosen, that is, migrate software viewers for multi-page TIFF format. That is to say, new viewers would be purchased for the computer platform or the viewer source code would be recompiled or reprogrammed for a new computer platform.

A *preservation action plan* is a plan for one or more preservation actions to be taken for the transfer of records to the archives, in accessioning the records, or for records being maintained. The Preservation Action Plan in Fig. 10 has a sequence of five preservation actions.

1. If the current computer platform does not have a multi-page TIFF viewer, or computer platform becomes obsolete, then acquire ulti-page TIFF viewer for the new platform.
2. Retrieve document images from case folder in the ECFS.
3. Reproduce the document images using the multi-page TIFF viewer.
4. Review the reproduced record to verify that the form and content are preserved.
5. If a record is reproducible and form and content are preserved, then store in the database that "on *current date* a viewer for TIFF multi-page format documents in the ECFS was acquired for *current platform*, and the viewer properly display the document images", else acquire new viewer for multi-page TIFF format.

Figure 10. Example of a Preservation Action Plan.

The first instruction triggers activity 1.2.4, *Acquire Capability to Apply Selected Preservation Method*. The second instruction triggers actions in activities A3.1, *Manage Information About Record*, and A3.2, *Manage Storage of Digital Components of Records*, to retrieve digital components for a specific series of records, and a specific class of records, i.e., in multi-page TIFF format. The third instruction triggers an action in A4.4, *Present the Record*, to use the presentation method (multi-page TIFF viewer) to reproduce the record. A person responsible for preservation performs the fourth instruction. The fifth instruction triggers an action in A3.1, *Manage Information about Record*, to store a record of the fact multi-page TIFF viewer was acquired and it was verified to reproduce the form and content of the document images in the ECFS.

Preservation actions are implemented using preservation methods. Preservation methods are software. Fig. 11 shows some examples of preservation methods. They include software for generic preservation methods such as integrity checks, methods for packaging or archiving many files as one, for refreshing media, for data base management, and for archival storage. They also include specific preservation methods,

for example, for reproducing records, for converting proprietary formats to standard formats, or for converting digital objects in proprietary formats to persistent objects.

Preservation Method Description	Examples of Corresponding Software
Check integrity of transferred records	Hash functions (MD5, SHA-1)
Package digital components for storage	TAR, WinZip, JAR
Storage Update Method	Tape Copy
Maintain information about records and digital components	DBMS (Oracle, Sybase)
Archival Storage	High Performance Storage System, DLT tapes
Reproduce records	TIFF and PDF viewers, X86 emulator
Update components	TIFFmaker, word2pdf, word2XML

Figure 11. Examples of Preservation Methods

The Baseline Requirements

One of the constraints (controls) on the Preservation Model is that it should satisfy the Baseline Requirements for Supporting the Production of Authentic Copies of Electronic Records developed by the Authenticity Task Force.[9] Baseline requirement 1, Controls over Records Transfer, Maintenance and Reproduction, is satisfied: (1) by activity A1.3.2 for creating *Terms and Conditions for Transfer*, (2) by activity A2.2 that compares the transfer with the *Terms and Conditions for Transfer*, (3) by activity A2.3.3 that takes the *Actions Needed to Preserve the Records*; and (4) by activity A4 that *Reproduces the Record* from maintained digital components.

Requirement 1a, unbroken custody of the record is maintained, is satisfied by institutional policies, and the Appraisal and *Bring In* (A2) activities. The access control and access privileges for activity A3.1, *Manage Information about Records*, and activity A3.2, *Manage Storage of Digital Components of Records*, satisfy requirement 1b, Security and control procedures are implemented and monitored. Requirement 1c, the content of the record remains unchanged after reproduction, is satisfied by selecting preservation methods that preserve content (activity 1.2.3) and verifying that records can be reproduced (activity 2.3.2).

Baseline Requirement 2, Documentation of the Reproduction Process and its effects, is satisfied by activity A1.2.3, *Selecting a Method to Apply to a Class of Preservation Objects*, and by activity A1.4, *Evaluation of Preservation*. Requirement 3, that the archival description for a body of records include information about changes to the records since they were first created, is satisfied by activity A3.3, *Update Digital Components*, and specifically by Preservation Action Plans that document the updates to digital components. It was concluded that each of the Requirements for Supporting the Production of Authentic Copies of Electronic Records is satisfied by some set of activities of the Preservation Model.

[9] Heather McNeil, InterPARES 1 Project, This proceedings.

Conclusion

The InterPARES preservation model provides a framework that archival institutions can use to manage the process of preserving authentic electronic records. Within that framework, a variety of preservation strategies can be developed that are dependent on the characteristics of the selected, transferred and accessioned records, institutional requirements, and the current and changing state of information technology. The preservation framework guides the development of preservation systems that can satisfy the Authenticity Task Force's Baseline Requirements for Supporting the Production of Authentic Copies of Electronic Records.

The walkthrough for a real case of electronic records selected for preservation shows that the model specifies how to develop the Terms and Conditions for Transfer, to assess whether a creator's records can be presumed authentic, to select preservation methods and to develop preservation plans. The model provides a framework for developing practical solutions to the preservation problem.

The walkthrough identified a number of refinements that were needed in version 5.1 of the preservation model. Some of these refinements were made in version 6 of the model. It was difficult to conduct the walkthrough without a data model for the kinds of information that are created, maintained and used in preserving electronic records. During InterPARES II, a data model will be constructed using the metadata that was identified during the walkthrough.

Additional walkthroughs will be conducted for case studies with different types of electronic records. This will ensure that the model can be realized in the real world for a variety of types of electronic records. It should also aid archivists in understanding how they can apply the model in their archival institutions.

Additional empirical research is needed in applying alternative preservation strategies to the same bodies of electronic records and determining their relative cost-effectiveness. This information would support the archival decisions as to the most cost-effective method to apply to a class of preservation objects.

While the walkthrough identified examples of the kinds of information that were needed to assess the authenticity of the electronic records in the Electronic Case Folder System, there was not actually enough information available in the case study to carry out the assessment. Experiments should be conducted to determine the kind of knowledge needed to perform the assessment, how to reason with degrees of belief, and whether the Benchmark Requirements and the method of assessment actually achieve what is intended.

When the method of assessment using the Benchmark Requirements results in a weak presumption of authenticity, the ATF prescribes that the preserver should attempt to verify the authenticity of the records. Research is needed in technical methods of authentication of preserved electronic records.[10]

---

[10] W. Underwood, A formal method for analyzing the authenticity properties of procedures for preserving digital records. *Proceedings of 2002 International Conference on Digital Archive Technologies*, (ICDAT2002) Academia Sinica, Taipei, Taiwan, pp. 53-64