Rivista quadrimestrale

Roberto CERRI

Elisa PANICUCCI

Elisa PANICUCCI

"ARCHIVI & COMPUTER"
Archivio Storico Comunale, Loggiati di San Domenico, 3 - 56027 San Miniato (Pi)
Tel. 0571 - 418381; Telefax 0571 - 406233; E-Mail: archilab@comune.san-miniato.pi.it
http://www.archi-lab.it

Abbonamento «persona fisica»:	Euro 31,00
Abbonamento «biblioteche, istituzioni, enti pubblici»:	Euro 36,10
Abbonamento estero:	Euro 51,60
Prezzo del singolo fascicolo:	Euro 12,90
Sconto per librerie:	10%
Sconto per soci Archilab:	30%

Le richieste di abbonamento devono pervenire alla **Carocci Editore** S.p.A., via Sardegna, 50 - 00187 Roma, tel. 06-42014260, fax 06-42747931, e-mail: riviste@carocci.it
I relativi importi devono essere versati sul conto corrente postale n. 77228005 intestato a Carocci Editore S.p.A., via Sardegna, 50 - 00187 Roma.

Per il momento è attivabile l'abbonamento 2002 alla sola versione cartacea. La redazione sta valutando l'opportunità di mantenere per il 2002 la versione elettronica sperimentata nel 2001. Nel caso che la versione elettronica non venga riattivata, sarà prevista una forma di rimborso per chi avesse già versato la relativa quota di abbonamento.

Sono disponibili i fascicoli e le annate 1991-2001 (un fascicolo Euro 12,90; un'annata Euro 51,60; sconto per librerie 10%; sconto per soci 30%). Le richieste devono pervenire ad Archilab, c/o Archivio Storico Comunale, Loggiati di San Domenico, 3 - 56027 San Miniato (Pi) - Tel. 0571 - 418381; Telefax 0571 - 406233; E-Mail: archilab@comune.san-miniato.pi.it e i relativi importi devono essere versati sul conto corrente postale n. 12252532 intestato ad Archilab, oppure effettuando un bonifico sul conto corrente bancario n. 10009004 intestato ad Archilab - Cassa di Risparmio di San Miniato - Ag. San Miniato - ABI 6300 - CAB 71150.

Autorizzazione del Tribunale di Pisa n. 8 del 23.03.1991.
Quadrimestrale - Spedizione in A.P. - 70% - Filiale di Pisa.
Editore: Archilab, Loggiati di S. Domenico, 3, 56027 San Miniato (Pi).
Stampa: Litografia Varo, via Carducci 51, 56010 Ghezzano (Pi).

II quadrimestre 2001 - Finito di stampare nel mese di gennaio 2002

ISSN 1121-2462

Archivi & Computer

Anno XI — Fascicolo 2/2001

ing Environment Group at the San Diego Supercomputer Center: Richard Marciano, Bertram Ludaescher, Ilya Zaslavsky, Amarnath Gupta, and Chaitan Baru.

References

[1] R. Moore, C. Baru, A. Rajasekar, B. Ludascher, R. Marciano, M. Wan, W. Schroeder, A. Gupta, *Collection-Based Persistent Digital Archives - Part 1*, «D-Lib Magazine», March 2000, http://www.dlib.org/

[2] R. Moore, C. Baru, A. Rajasekar, B. Ludascher, R. Marciano, M. Wan, W. Schroeder, A. Gupta, *Collection-Based Persistent Digital Archives - Part 2*, «D-Lib Magazine», April 2000, http://www.dlib.org/

Kenneth Thibodeau

Digital Preservation Techniques: Evaluating the Options

L'autore parla dei fattori che rendono problematico preservare a lungo termine i documenti elettronici e gli oggetti digitali e analizza i metodi di conservazione raggruppandoli in tre categorie a seconda del tipo d'intervento operato: mantenimento del formato originario dei documenti e della relativa tecnologia, conversione del formato originario dei dati in un formato compatibile con la più recente tecnologia, trasformazione del formato originario in un formato indipendente da specifiche tecnologie e quindi non minacciato dall'obsolescenza di software e hardware provocata dal continuo progresso tecnologico. Il contributo analizza vantaggi e svantaggi dei tre tipi di soluzioni che, comunque, non sono rigidamente alternative ma possono essere integrate.

As we face a future where digital information appears certain to become more and more prominent in the conduct of affairs in government and in business, there is a major element of uncertainty about the survival of information in digital form. As a group of prominent experts reported to the President of the United States in 1999, «Our Nation's security, commerce, education, and well-being depend increasingly on our information infrastructure. It is thus critical to ensure the survivability of that infrastructure.... Survivability includes long term preservation of information...»[1]. But today, no one is in a position to guarantee the survival of most forms of digital information.

Of the many problems we face in trying to preserve digital information, the two that are best known are media fragility and technology obsolescence[2]. The reverse side of obsolescence is technological progress. It is not only the cause of obsolescence, but also compounds the problem of preservation: besides digital objects becoming obsolete, new types and formats of digital information are be-

[1] President's Information Technology Advisory Committee.

[2] International Council on Archives. Committee on Electronic Records. *Guide for managing electronic records from an archival perspective. Studies - Études 8. February 1997*, Paris, ICA, 1997.

ing created constantly. There are other problems as well. A basic problem derives from the fact that there are many elements of indeterminacy in the digital information objects we want to preserve.

Consider even a relatively simple example: that of a document which contains nothing but natural language text. At a basic, empirical level, it is impossible to store the document in digital form. Operating in character mode, it would be possible to store digital representations of the characters that make up the content of the text, along with other digital codes that indicate how the text should be organized and how it should appear when presented to humans. Alternatively, one could store a digital representation of a printed or viewable image of the document. But in either case, in order to retrieve the document, it is necessary to process the stored binary digits through software running on hardware. The hardware and software that is used can have as much effect on the document that is output as the stored bits themselves. Obviously, for example, the same document will look different on two different computers with different size monitors and screens with different resolutions. Even on the same computer, differences in the appearance of a document can occur by changing the software settings for resolution or number of colors displayed. The presentation of a file can be altered because of state changes when different software is run concurrently on the same computer.

The particular software used to render a document can change its structure and appearance without any change in the content of the digital file in which the document is «stored.» Attributes of digital documents can also change as a result of simple user actions which do not change any of the digital content of the document. Many application software products give users several ways to change the appearance of a document; such as by selecting different zoom ratios, or choosing between print image and so-called 'draft' or 'normal' viewing modes. Users can set default values on their computers that cause documents to appear differently than their authors intended. For example, it is possible to change the default background color on a web browser so that documents which were created, and stored, with a white background appear on a green or grey background. Software can change the structure and appearance of a document in ways that are not intended or controlled by the user. An example is what happens when the user simply changes the size of the window in which a document is displayed.

In the case of an HTML document, when window size is reduced from full screen, Netscape Explorer will preserve the paragraph structure, the size of type, and the length of lines relative to the width of the window in which the page is viewed; however, it achieves this at the expense of the line flow. The same paragraph will have different numbers of lines, with different words on each line in two differently sized windows. If the same file is opened in Corel WordPerfect, changing window size produces different effects. WordPerfect retains paragraph structure, relative length of lines and the words appearing on each line. It conserves these attributes by changing the size of the type displayed. Microsoft

Word handles window changes differently, as shown in Figure 4. Word holds type size, as well as line and paragraph flow, invariant, with the result that users have to scroll horizontally back and forth across the window to read entire lines.

Another major factor contributing to the indeterminacy of digital objects is the fact that software is imperfect, and may behave in unpredictable and inexplicable ways. Such problems can cause systems to fail. An example is the «General Protection Fault» which occurs on Windows systems. A web site entitled, «The Internet Help Desk», states: «A general protection fault (GPF) occurs when Windows attempts to access a memory space previously allocated to another program. Any code present in the memory space at the time of the illegal access risks data corruption. Generally, only data held temporarily in memory at the time of the attempted access becomes corrupted, but program files can become permanently damaged as well. .... Remember that it may be difficult or impossible to treat the cause of GPFs. Because there are so many potential variables involved in troubleshooting GPFs, it is sometimes only possible to treat the symptoms and not the root cause» (http://w3.one.net/~alward/gpf.html). The problem is not limited to conflicts between competing applications. The online «Microsoft Knowledge Base» describes many problems with Microsoft products in terms of 'symptoms' and 'work arounds', but does not identify their causes or solutions (http://search.support.microsoft.com/kb/c.asp?ln=en-us&sd=gn).

These examples illustrate that the structure and appearance of digital documents can be changed by a variety of factors, including differences in hardware and in software settings, vagaries of software and machine states, and user options. The complexity of determining exactly what a digital document is only increases as one moves beyond the simple case of natural language text used in the preceding examples towards more complex types of digital objects, such as text with embedded images, spreadsheets, databases, etc.

Given the variety of factors that can introduce elements of indeterminacy in digital documents, arguably it is impossible to determine exactly what a digital document is except on a specified system at a given moment in time. This challenges the very notion of preserving a digital document. In the simplest sense, a document can be said to be preserved if and only if it is possible to retrieve the document at some point in time and know that it is the same as it was at some prior time. But this simple concept of preservation is practically without meaning in a digital environment. As the above examples show, one cannot know how a digital document appeared at any time in the past unless one saw it at that time. Moreover, the problem is not necessarily one of time. Two users could open the same document, stored on a server to which they both have access, at the same time with notable differences between the two renderings resulting from any, or any combination, of the factors cited above. If these two users could be said to be looking at the same document, or if a single user can be said to be still viewing the same document after changing the size of the window in which it is presented, then the integrity of the document does not necessarily involve all of its properties. Following this line of reasoning, a document would be said to be preserved over time if all of its essential properties remained unchanged, even though other properties either changed or were indeterminate.

But this recognition raises the question of what are the essential properties of documents? There are no ready, widely-accepted, or well-established answers to this question. The InterPARES project is an international collaboration which is trying to define the essential properties of one class of digital documents, electronic records[3]. But this is a work in progress, and even when it achieves its objective, the results will not necessarily be applicable to other documents which are not records. This consideration reveals another aspect of the problem of digital preservation: the requirements for preservation are not well articulated or established. Much of the published literature on digital preservation simply ignores this problem, proceeding on the at least tacit assumption that digital preservation entails preserving whatever technological artifacts are produced using computers.

In some cases, this assumption is justified, but it must be recognized that there is a spectrum of options for digital preservation ranging from an accent on preserving the technology on the one end to preserving the things produced with the technology on the other. Where preservation solutions should fall on this spectrum depends to some extent on what is to be preserved. Solutions for things, such as computer games, that are essentially embodiments of the technology should fall closer to the technology end of the spectrum. But in many cases it is the properties of the thing itself, rather than the technology used to produce or store it that are important. For example, in preserving a digital photograph what is important over time is that the photograph continue to look as it did originally. If it were necessary to replace all of the technology originally used to produce or store the photograph in order to maintain the appearance unchanged, it would be legitimate to do so. In fact, it could be argued that the photograph would be best preserved as such by outputting it to a stable film or photographic paper base rather than risking the uncertainties entailed by continuing change in digital technology.

The preservation of records, too, should emphasize their properties as records, rather than technological characteristics that are merely coincidences of the technology used to create or store them. In preserving electronic records, it is necessary to take into account that they are objects which have properties that derive from three different domains. All records have characteristics that derive from the information technology used to create and keep them. All records also have properties that relate to their documentary form. Finally, all records have properties that derive from their archival nature, specifically from their provenance and archival bond. What is important is the record's ability to stand for the acts or facts to which it attests, rather than that it embodies certain digital formats. Of course there are cases of electronic records where technological characteristics are essential, especially where there is no way to capture the record except in digital form. A virtual engineering drawing of an airplane that permits 3-dimensional rotation is such a case. However, even in such a case, it remains necessary to preserve those attributes that make the object a record. Otherwise, one is left with preserving something that is merely a curiosity, not a record.

The final complicating factor that makes digital preservation challenging comes into view when one considers the purpose of preservation. Information is preserved on the presumption that it has value over time. The value of information is realized in use. The possibility of making use of information depends on the possibility of discovering and delivering it. We can assume that continuing progress in digital information technology will produce better and better tools for finding information of interest and for delivering it in a useful form. Digital preservation strategies should aim at making it possible for future researchers to discover and obtain what they are looking for. There is an inherent tension between the objective of taking advantage of future improvements in information technology and any requirement to preserve characteristics which are dependent on current or past technology.

What are the possibilities for addressing the multi-faceted problems involved in digital preservation? There are quite a number of approaches that have been suggested and discussed, but very few that have actually been tried. Fewer still have proven their mettle in actual use. Methods for preserving electronic records, and digital objects, that have been used, or proposed, to date can be arranged in three major groups ranging from those most tightly bound to specific technology to those which are most independent of given technology. These three groups can be labeled maintenance of technology, data format conversion, and transformation to persistent object form.

Maintenance of technology includes those methods which aim at retaining things produced with the technology in their original formats and retrieving, processing and delivering them either by keeping the original hardware and software in operation or by imitating the necessary technology. Data format conversion abandons the original hardware and software and overcomes obsolescence by reformatting data files to newer formats that can be retrieved and used with current software. Transformation to persistent form creates a proxy or replica for digital objects in a format that is relatively independent of specific technology and, therefore, impervious to continuing changes in technology. These categories are not necessarily mutually exclusive. Some approaches that have been suggested combine elements of more than one of the three groups. Each admits of some variation.

Maintaining technology

Within the approach of maintaining technology one can distinguish two subtypes. One, which tries to keep the necessary hardware and software in operation,

[3] A.J. Gilliland-Swetland and P.B. Eppard, *Preserving the Authenticity of Contingent Digital Objects: The InterPARES Project*, «D-Lib Magazine», July/August 2000, volume 6, number 7/8.

can be named «*maintaining the original technology.*» The second approach could be characterized as «*imitating the original technology.*» This approach does not strive to keep all of the original technology working, but tries either to simulate the old technology on new systems or to run the original software «in emulation» on future hardware.

Maintaining technology appears to have obvious advantages deriving from the fact that documents are retained in their original formats. It is not unreasonable to assert that a document which is retained in its original format and accessed using either the original technology or replacement technology which in some way imitates the original is preserved intact. However, by the same measure this approach perpetuates the problems of current technology, including the incertainty about the exact presentation of the document, imperfections in software, and proprietary elements that can only be assumed, but not guaranteed, to remain operable in a consistent fashion across generations of technological change. Perhaps more serious, maintaining technology will become increasingly complex over time. Given that hardware, operating systems and applications software change at frequencies counted in months, over a few decades, this approach will entail maintaining thousands of combinations of technological components in functional order. The combinatorics become even more intimidating as one broadens the area of concern from individual files to systems, distributed applications, Internet applications, software mediators, metacomputing, and information grids. Even it proved to be possible to get the thousands and thousands of combinations of software and hardware components to work properly, the possibility of providing the necessary technical support – including end user support in a context where the only valid assumption is that most users will have never encountered most of the technology – absolutely defies credibility. Consideration of users raises another difficulty with this approach: a commitment to obsolete formats and obsolete technology runs directly counter to the desirability, and even the possibility of using the best available technology to deliver service. If it is necessary to use obsolete technology to access records, then it will not be possible to take advantage of improvements in technology over time.

If we turn from general considerations to the preservation of archival records, preservation strategies that rely on maintaining technology apparently reduce problems of integrity and authenticity of the records because their binary storage formats are kept unchanged. However, if access to the records involves using the software that was used to create the records in the first place, in many cases the integrity of the records will be questionable whenever they are used because the same software used to create them can be used to alter them. If we could be certain that the record that is rendered when a stored file is retrieved and processed through the software is invariant over time, we could assume the record is authentic; however, the imperfections that exist within software products, the unpredictability of what happens when different software products interact, and the variability in presentation discussed above give the lie to this assumption. Maintaining technology is not a basis for asserting the authenticity of records over time simply because it never gets to records; that is, it never actually addresses the essential properties of records as such. Rather it remains fixated on technology.

Data Format Conversion

Data Format Conversion involves reformatting digital files when the software necessary to process these files becomes obsolete. It abandons both original hardware and software but moves the files being preserved forward to keep pace with technological change One can distinguish two basic strategies for converting data formats. One, which might be called '*versioning*', relies on a linear, one-to-one conversion from an older to a newer version of the 'same' format (e.g. a Lotus «wk3» spreadsheet to a «wk4» spreadsheet). The need for such conversions usually arises when a software producer changes a software product in such a way that the files created using the changed software have properties that were not present or were encoded differently in earlier versions. When this happens, the producer usually provides the capability for converting files from earlier formats to the current one.

The other format conversion strategy, which we could call '*format standardization*,' involves changing the files to standard formats. Open, proprietary, or ad hoc standards may be used. Examples include changing word processing files to plain ASCII or into Adobe Corporation's 'pdf' format, or converting data files containing nested into basic relational database format by normalizing the data.

Format conversion methods have an advantage over methods that seek to maintain technology in that, at any given time, a much narrower variety of hardware and software will need to be maintained. However, both versioning and format standardization are likely to be relatively short term solutions because both proprietary and standard formats become obsolete. Format conversion, especially versioning, is inherently short-term and ad hoc; moreover, there is no guarantee that there will be a replacement format in either case. In addition, newer formats may not be equivalent to older ones. New products often include new features and new functionality that may change the character of the objects being retained over time. Formats produced by commercially available products may conform to published standards, but they often include features that go beyond or are not addressed in standards. These features are likely to be lost in conversion to standard form. In sum, there is a risk of alteration at every conversion. Such alterations, especially in the case of proprietary formats, may be unknown and uncontrolled and uncontrollable. Finally, as with maintaining technology, albeit to a lesser degree, the complexity of this solution increases as a function of the number of formats maintained

Considering the specific case of the preservation of records, we can see that format conversion does not explicitly address any of the attributes of records;

therefore, it cannot be said to be suited to the preservation of records in any meaningful way. This method addresses problems of technology obsolescence. It does not get to requirements for preserving records. The integrity of a record under repeated conversions over time may be impossible to demonstrate. Moreover, format conversions coupled with replacement of software may entail difficulties with the faithful reproduction of a record because the different combination may produce lost, added or altered features or functionality.

Transformation to Persistent Object Form

Transformation to Persistent Object Form defines an abstract model, or models, for each digital object that is to be preserved. The basic model for all objects should identify all the significant components contained in the object and articulate the structure which relates the components to each other and to the whole. The transformation consists of applying the model to the object by marking its content with tags that identify where each component occurs in the object. The model and the related tags should be expressed in a simple, coherent form that is not dependent on any specific hardware or software. In this manner, the digital objects become self-describing. This approach has been developed at the San Diego Supercomputer Center (SDSC) and is being pursued in a number of research projects related to archives of electronic records and other collections, such as scientific data sets. The work currently uses the eXtensible Markup Language (XML) for the syntax of persistent object transforms. (www.sdsc.edu/NARA and www.npaci.edu).

In the process of transformation, any proprietary markings likely to be subject to rapid obsolescence are replaced by open tags. However, many native formats include special codes that do not identify what the content represents, but only how it is to be presented. Where the appearance should be preserved, it is possible to represent it in a persistent form though style sheets constructed in eXtensible Style Language (XSL).

This process can be applied to collections or sets of digital objects as well as to individual objects. Given that each of the members of a set is characterized by its own model, the model of a set need only represent the structure of the set and the placement of members in that structure.

Transformation to persistent object form renders the objects being preserved relatively independent from the information technology infrastructure used to materialize the objects at any given time. Infrastructure independence operates both retrospectively and prospectively. Retrospectively, the transformation eliminates the dependency of digital objects on the technologies originally used to create or store them. Prospectively, it makes it possible to deliver the objects using future technologies, but retaining their original properties. This can be accomplished though middleware that translates between the persistent object form and the formats used by the target technologies. SDSC is developing such middleware called Mediation of Information using XML (MIX). As technologies change over time, the persistent object form can remain stable. The only thing that needs to be changed is the translator.

This approach is obviously advantageous in the case of collections which are inactive over long periods of time. In a persistent archives, the collections are not materialized as such. Thus they are relatively immune to repeated changes in technology over time. What is retained in the persistent archives is all of the persistent objects, appropriately tagged and encapsulated in metadata, and the models of the collections or sets. When a set is to be used, its model is translated into a format that can be implemented in the target technology and used to rebuild the collection structure. Then the objects are placed in their appropriate positions in that structure.

Persistent object transformation aims at independence of technological infrastructure. It reduce threats to integrity and authenticity by minimizing changes over time. At the same time, it embeds changes in a comprehensive information management architecture designed explicitly for long-term preservation. The method is very extensible. It has been applied in prototypes involving databases reflecting 30 years of database management technologies, a collection of one million e-mail messages, a collection of 10,000 digital images of art works and other objects held by museums along with databases about both the images and the original art and objects, geographical information systems, and others. This method facilitates use of future, advanced technologies, without requiring any change in what is preserved[4].

If we focus on the application of persistent object transformation specifically to records what distinguishes it from other preservation methods is, above all, that it explicitly addresses archival requirements for the preservation of authentic records. Given that it is applicable to collections with arbitrary and arbitrarily complex hierarchical structures, it preserves archival bonds and archival fonds as well as individual records. Minimizing changes over time reduces threats to integrity and authenticity. Requirements for integrity and authenticity are controlled explicitly and specifically. Furthermore, persistent object preservation provides for the performance of archival functions as well as the persistence of the records. It is designed so that use of advanced technologies for search, retrieval and delivery does not compromise integrity[5].

[4] R. Moore, C. Baru, A. Rajasekar, B. Ludascher, R. Marciano, M. Wan, W. Schroeder, A. Gupta, *Collection-Based Persistent Digital Archives*, «D-Lib Magazine», March and April 2000 (http://www.dlib.org/dlib/march00/moore/03moore-pt1.html and http://www.dlib.org/dlib/april00/moore/04moore-pt2.html)

[5] K. Thibodeau, *Building the Archives of the Future: Advances in Preserving Electronic Records at the National Archives and Records Administration*, «D-Lib Magazine», February 2001 (http://www.dlib.org/dlib/february01/thibodeau/02thibodeau.html)