Autorità per l'informatica nella Pubblica Amministrazione

La conservazione dei documenti informatici - Aspetti organizzativi e tecnici

Seminario di studio

Roma 30 ottobre 2000

Programma

I SESSIONE: LA FUNZIONE CONSERVATIVA. PROBLEMI, STRATEGIE, ESPERIENZE,

9:00   **Apertura del seminario**

Salvatore Italia Direttore Ufficio Centrale per i beni archivistici

Guido Mario Rey –Presidente dell'Autorità

9:15   **Relazione introduttiva**

Carlo Batini (Autorità per l'Informatica nella Pubblica Amministrazione)

9:35   **La gestione e la conservazione dei documenti informatici nelle pubbliche amministrazioni italiane (The management and the preservation of electronic records in the Italian public administrations)**

Antonio Massari (Autorità per l'Informatica),

10:10   **Information Management Architecture for Persistent Object Preservation**

Reagan W. Moore (University of California)

10:40   **Intervallo**

11:00   **The strategies for the preservation. Projects and perspectives in the US Federal Government**

Ken Thibodeau (National Archives USA)

11:30   **The Swedish programs for the electronic records preservation and access**

Kristiansson Gorlan (Riksarkivet, Sweden)

12:00   **Responding to the challenges and opportunities of ITC: the new record manager**

Seamus Ross (University of Glasgow)

12:30   **Discussione**

13:00   **Intervallo per il pranzo**

II SESSIONE. I METODI PER LA CONSERVAZIONE, IL RUOLO DEI METADATI, LE POTENZIALITÀ DI XML

14:00   **Relazione introduttiva**

Maria Guercio (Università di Urbino)

14:30   **A logical model for the electronic records authentication**

Bill Underwood (Georgia Tech Institute)

15:00   **Metadata: Archival Concept or IT Domain?**

Peter Horsman (IT Committee, International Council on Archives, president)

15:30   **The potential of markup languages to support descriptive access to electronic records: the EAD standard**

Anne Swetland Gilliland (University of Los Angeles)

16:00   **XML standards for business-to-business and business to government communication**

Zachary Coffin (eXtensible Business Reporting Language)

16:30   **XML, uno standard per gli archivi informatici (XML, a standard for the electronic records)**

Daniele Tatti (Autorità per l'informatica)

Indice

Premessa

Il seminario, organizzato dall'Autorità e dall'Ufficio Centrale per i Beni Archivistici del Ministero per i Beni e le Attività Culturali, ha la finalità di affrontare la questione cruciale della conservazione a lungo termine dei documenti informatici, da un lato sensibilizzando le amministrazioni pubbliche che hanno già avviato o programmato processi di automazione dei sistemi documentari, sia di avviare - anche attraverso un confronto internazionale - l'analisi e la verifica delle modalità e degli strumenti che consentono il mantenimento nel tempo dell'integrità e dell'autenticità delle memorie digitali nonché la loro accessibilità. Il tema ha acquistato attualità e rilevanza crescenti in relazione alla larga diffusione di tecnologie dell'informazione e della comunicazione per la gestione dei flussi amministrativi e documentari. Richiede , tuttavia, la soluzione di numerosi problemi di natura organizzativa e tecnica che gruppi internazionali di ricerca e istituzioni nazionali di tutto il mondo hanno cominciato a studiare da qualche anno con impegno crescente soprattutto allo scopo di individuare soluzioni scalabili e di basso costo che rendano possibile anche alle organizzazioni di piccole e medie dimensioni avviare interventi generali di automazione dei processi amministrativi e delle attività di trasmissione e tenuta dei documenti.

In particolare, **nella prima sessione**, il seminario si propone di offrire alle amministrazioni pubbliche italiane - soprattutto ai responsabili dei sistemi informativi e ai responsabili dei sistemi documentari - un panorama della specifica situazione nazionale e dei progetti e delle esperienze operative più interessanti condotte in altri Paesi (Stati Uniti e Svezia) in stretta collaborazione con più ampi gruppi di ricerca internazionali, insieme a un'analisi delle esigenze di formazione e riqualificazione dei profili professionali destinati ad assumere il nuovo difficile ruolo di conservatore delle memorie digitali.

La seconda sessione è destinata ad esaminare le questioni più strettamente tecniche della conservazione, in particolare i diversi metodi oggi disponibili per affrontare l'obsolescenza tecnologica e, in special modo, le potenzialità dei linguaggi di marcatura (SGML, XML) per il trattamento e la gestione dei metadati che garantiscono la tenuta a lungo termine del patrimonio documentario e lo sviluppo di standard specifici (Encoded Archival Description).

Preservation and migration of electronic records: the state of the issue[13]

Kenneth Thibodeau[14]

The problem of preserving electronic records

The two-edged sword of continuing progress and rapid obsolescence of information technology is the most often cited, but perhaps not the most significant challenge archives face in the endeavor to preserve electronic records. Organizations rely more and more on digital technology to produce, process, store, communicate, and use information in their activities. Thus, the quantity of records being created in electronic form increases. In the experience of the National Archives and Records Administration of the United States, it increases exponentially. The technological challenge is compounded by the continuing extension of information technology in terms of the types of information objects it produces, and again in terms of its applicability to different spheres of activity and different types of actions within those spheres. The resultant records are increasingly diverse and complex. The impact is not only on individual records, but on the archival fonds as a structured whole.

Approaches to the problem of preserving electronic records

The field of information technology has, by and large, ignored the problems of long term preservation. If anything, one could say that the market has tended to exacerbate the problem of preserving electronic records. The pressures of competition have led the industry to obey Moore=s law, replacing both hardware and software on a frequency of two years or less.

In one area, however, there has been some improvement in recent years: that of digital storage media. From the 1980s there was a trend towards storage media that were more fragile and less stable over time. In recent years, this trend, if not reversed, has been offset somewhat by the introduction of more stable and reliable media.[15] Current research and development efforts offer the prospects of improved options for long term storage of digital information, notably in the areas of ion-milling and holographic media. But archival concern with digital media should not be limited to their durability. The ICA *Guide to Managing Electronic Records* sets out seven criteria for media used for preserving electronic records:

$ open standards for digital recording on the medium,
$ robust methods for preventing, detecting and reporting errors,
$ sufficient market penetration,
$ known longevity,
$ known susceptibility to degradation or deterioration,
$ a favorable cost/benefit ratio, and
$ availability of methods for recovering from loss.[16]

Whatever relief archives may find in the area of digital storage is more than offset by the increasing diversity, complexity and spread of electronic records. In recent years, increasing attention has been devoted to problems of digital preservation in a variety of spheres and professions. Several

---

[13] This paper was presented at the XIVth International Congress on Archives, Seville, Spain, September 22, 2000. The views expressed are the author's and not necessarily those of the National Archives and Records Administration.

[14] The author is Director of the Electronic Records Archives Program, National Archives and Records Administration, U.S.

[15] Charles M. Dollar. Authentic electronic records: strategies for long-term access. Chicago: Cohasset. 1999. Pp. 58-60.

[16] International Council on Archives. Committee on Electronic Records. Guide for managing electronic records from an archival perspective. Paris. 1997.

different approaches has been proposed. A few have been tried in test mode, fewer in actual practice. In practice, the experience of archives is largely limited to relatively simple technical formats, such as flat files. Some institutions have developed computer applications for preserving potentially complex databases. These include CONSTANCE at the National Archives of France, AERIC at NARA, ERICSON at the National Archives of Canada, and similar systems in Sweden, the United Kingdom and elsewhere. Significant preservation projects addressing the actual preservation of digital formats, at various stages of research or development, include the bundles proposal of the British Standards Institute,[17] the CEDARs project at the University of Leeds, England,[18] the Victoria Electronic Records System in Australia,[19] the emulation experiment at the Royal Library in The Netherlands,[20] the Universal Preservation Format sponsored by the WGBH Educational Foundation in Boston,[21] and the Highly Integrated Information Processing and Storage technology being developed at Carnegie-Mellon University in the U.S.[22]

Current initiatives are pursuing quite a variety of approaches. The proposed solutions can be categorized into five broad categories:

$ preserving the original technology used to create or store the records;
$ emulating the original technology on new platforms;
$ migrating the software necessary to retrieve, deliver, and use the records;
$ migrating the records to up-to-date formats; and
$ converting records to standard forms.

These approaches define a spectrum ranging, in broad terms, from no change in the records or the technological context in which they exist to one in which the original hardware and software have disappeared and the digital format of the records has changed. Each of these methods has pros and cons. None of them is entirely satisfactory. On the one hand, in general, one can say that the closer one stays to the original technology and original digital format of the records, the less the problem of authenticity; however, it is also obvious that the closer one stays to original technology, the more complex and more impractical the approach becomes over time. More complex because, as records continue to accumulate over time, there will be more and more varieties of technology that the archives would have to maintain. More impractical because, first, support for obsolete technologies will eventually disappear and, second, the distance and difference between the preserved technology or technical artifacts B including the records B and the best available technology for preserving, managing, retrieving and delivering the records will increase continuously. On the other hand, while moving ahead as technology progresses can eliminate such practical problems, it can entail loss or corruption of records.

<u>The need for an archival approach to preserving electronic records</u>

All of these approaches to preserving electronic records have in common the objective of solving technological problems related to the passage of time. None of them actually focus on the objective of preserving records. This technological orientation is misdirected because success in solving

---

[17]British Standards Institution. Bundles for the perpetual preservation of electronic documents and associated objects. Public Draft for Comment - IDT/1/4: 99/621800DC.London. 1999.

[18]David Holdsworth and Derek M. Sergeant. A blueprint for representation information in the OAIS mode. In: Eighth Goddard Conference on Mass Storage Systems and Technologies, B. Kobler and P.C. Hariharan, editors. Maryland, Goddard Space Flight Center, 2000. Pp. 413-28.

[19]Public Record Office Victoria. Victorian Electronic Records Strategy. Final Report. 2000. <http://www.prov.vic.gov.au/vers/final/finaltoc.htm>

[20]Jeff Rothenberg. An experiment in using emulation to preserve digital publications. Den Haag. Koninklijke Bibliotheek. 2000.

[21]Dave MacCarn, Toward a universal data format for the preservation of media. SMPTE Journal, July 1997 v106 n7 p477-479. See also <Http://info.wgbh.org/upf/>

[22]http://www.ece.cmu.edu/research/chips/

technological problems does not necessarily imply any success, or even relevance, in addressing archival requirements for the preservation of records.

Logically, archival principles and objectives should dictate the requirements that technical solutions must satisfy. Archival requirements for preservation must be based on the conception of electronic records, not as the products of computer applications, but as the instruments and by-products of the practical activity of a records creator. The ultimate criterion for success in the preservation of electronic records is not whether they remain true to some given technological materialization, but whether they continue to provide authentic evidence of the activities in which they were created.

<u>An architecture for archival preservation</u>

Clearly, the archival profession needs to determine specific requirements for the preservation of different types of records, and also to guarantee respect for provenance and the integrity of archival fonds over time. The InterPARES project, directed by Professor Duranti, brings together archivists from universities and archival institutions, along with computer and information scientists and engineers, from around the world in a concerted effort to delineate specific archival requirements for preserving authentic electronic records. InterPARES is working to define the archival requirements for authenticity on the basis of archival science and diplomatics.[23]

Simultaneously, the InterPARES Preservation Task Force is examining technical issues related to digital preservation and developing a formal model of the preservation function as viewed from the perspective of the juridical or physical person responsible for preserving electronic records. While this work is still in progress, there are several ideas which have been proposed that are worth citing at this time. One key idea is that, strictly speaking, it is not possible to preserve electronic records; it is only possible to maintain the ability to reproduce electronic records. It is always necessary to retrieve from storage the binary digits that make up the record and process them through some software for delivery or presentation. B Analogously, a musical score does not actually store music. It stores a symbolic notation which, when processed by a musician on a suitable instrument, can produce music. B Presuming the process is the right process and it is executed correctly, it is the output of such processing that is the record, not the stored bits that are subject to processing.[24] This concept has important consequences. It shifts priority in preservation of electronic records from their storage over time, to the integral processes of putting the records into archival storage, getting them out of storage, and delivering them to future researchers. The recognition that electronic records must inevitably be reproduced accentuates the importance of being able to demonstrate the integrity and authenticity of the records. This entails extending the traditional concept of an unbroken chain of custody into one of an unbroken process of preservation. As defined in the ICA Guide, AAn electronic record is preserved if and only if it continues to exist in a form that allows it to be retrieved, and, once retrieved, provides reliable and authentic evidence of the activity which produced the record.@[25] Demonstrating the authenticity of electronic records depends on verifying that:

1. the right data was put into storage properly;
2. either nothing happened in storage to change this data or alternatively any changes in the data over time are insignificant;
3. all the right data and only the right data was retrieved from storage;
4. the retrieved data was subjected to an appropriate process, and
5. the processing was executed correctly to output an authentic reproduction of the record.

Parallel to the InterPARES project, the National Archives and Records Administration is sponsoring research into the development of an information management architecture designed to address archival

---

[23] Anne J. Gilliland-Swetland and Philip B. Eppard. Preserving the authenticity of contingent digital objects. The InterPARES project. D-Lib magazine. July-August 2000.

[24] Preliminary report from the chair of the Preservation Task Force to the Director of the InterPARES project, March 30, 2000.

[25] ICA. Guide. P. 35.

requirements for the preservation of electronic records. This architecture implements the proposed ISO standard for an Open Archival Information System (OAIS).[26] The architecture extends that general reference model by articulating archival requirements. To address the basic problem of continuing change in technology over time, the architecture postulates that archival information systems should independent of the particular technology used to implement them at any time. That is, an archival information system should be built in such a way that it is possible to replace any component of hardware or software used in the system with minimal impact on the rest of the system and with no impact on the preserved collections of records.[27]

<u>Collection-based persistent object preservation</u>
The information management architecture is being developed in the U.S. National Partnership for Advanced Computational Infrastructure. The Partnership is a collaboration of 46 institutions nationwide, and 6 foreign affiliates, with the San Diego Supercomputer Center serving as the leading edge technical resource. The research is addressing archival requirements for preservation of records, including respect for provenance. Rather than focus on technological problems, the method focuses on the objects that are to be preserved. In this case, the objects are records and also collections of records, as organized within archival fonds at all levels of hierarchy.

The method of collection-based persistent object preservation consists of identifying the properties of the objects to be preserved; expressing those properties in explicit, abstract models; and applying those models to transform the objects into an independent technological format suitable for long-term preservation. In the archival domain the development of this method started with the conception of the essential properties of records expressed in the ICA Guide on electronic records; that is, *AA record is recorded information produced or received in ... an institutional or individual activity and that comprises content, context and structure sufficient to provide evidence of the activity regardless of the form or medium.@[28]* The essential *structure* of a record is its documentary form. This form may be expressed in the digital format in which the record is stored, but it is not necessarily identical to the digital format. Therefore, a transformation of the record which replaces one digital method with another one that is more suitable to long-term retention, preserves the record so long as it maintains the essential documentary form of the record. The immediate *context* of a record is its archival bond: the position of a record with respect to other records in the archival fonds. In our research, we have extended the list of essential properties of records beyond content, structure and context to include the *appearance* of the record. We are also addressing a special type of content that is unique to electronic records: hyperlinks.

Persistent Object Preservation expresses the structure of records using eXtensible Markup Language (XML) Document Type Definitions. The method encapsulates records using the metadata defined in these models, transforming records into a format that is independent of any specific technology. The research has demonstrated that this method can be applied to collections of records as well as to individual records. That is, one can construct a Document Type Definition to capture and preserve the structure of any archival collection, of arbitrary complexity, from individual files through series and classes to entire archival fonds.

The research is exploring different ways of preserving the appearance of records. One way is to use a technology known as Multi-Valent Documents to capture and retain a bitmapped image of the document. MVD enables the image to be retained not as a version of the document, but as a layer of the

---

[26]Consultative Committee on Space Data Systems. Reference Model for an Open Archival Information System (OAIS). Draft Recommendation for Space Data System Standards, CCSDS 650.0-R-1. Red Book. Issue 1. May 1999. <http://ccsds.org/RP9905/RP9905.html>

[27]Kenneth Thibodeau, Reagan Moore, and Chaitanya Baru. Persistent object preservation: advanced computing infrastructure for Digital Preservation. European commission. Proceedings of the DLM-Forum on electronic records. European citizens and electronic information: the memory of the information society. Brussels, 18-19 October 1999. Luxembourg. Office for Official Publications of the European Commission. 2000. Pp. 113-118.

[28]ICA. Guide. P. 22.

document object modeled as an acyclic directed tree.[29]   Another possible means of preserving appearance is through the eXtensible Style Sheet Language (XSSL) available in the XML standard. Using style sheets to capture the attributes of appearance is especially advantageous for types of applications, such as databases and geographic information systems, where stored data elements may participate in many different records. In such systems the records are likely to be expressed as views, forms, or reports which extract specific subsets of the data and present them in predefined formats. A different style sheet can be defined for each of these formats

The method extends beyond the preservation of archival collections of records over time. It also addresses the key archival functions; notably, the accessioning of records into the archival repository, the establishment of intellectual control over the records, and the delivery or dissemination of the records to researchers. This extension of the persistent object approach is consistent with the basic premise of object oriented methodology which starts with the recognition that an object has behaviors or methods, as well as attributes. One of the essential behaviors of a record is that it occupies a specific position in relation to other records in the archival fonds. This behavior expresses the immediate context of the record and is the basis for arriving at its significant context; that is, the activity of which the record provides evidence.[30] The transformation of records into a persistent object format not only enables the records to be preserved indefinitely into the future, it also makes it possible to benefit from advanced technologies, which have not even been invented yet, to search, access and deliver the records in the future. This is made possible though the separation of context, structure and appearance in explicit schemas expressed in simple textual form. Over time, it will not be neccessary to migrate the materials stored in persistent object form to new technologies, but only to interpret the schema metadata so that it can be used in future technologies.

Viability of the persistent object preservation method

The initiative to develop the collection-based persistent object method for preserving electronic records is still in the stage of research and development, and will remain in this stage for some time. Nonetheless, there are substantial reasons, in both the technical and the archival domains, to assume that it will be successful. In the domain of technology, two facts should be highlighted. First, the research is not developing any special technologies to suit archival needs. Rather, it is building archival solutions on the basis of technologies which are seen as essential to the next generation Internet and information infrastructure and as keys to electronic commerce and electronic government. Archives should benefit, therefore, from widespread market support for the enabling technologies. Second, while the research addresses archival requirements specifically, the method has broad application in other areas, such as digital libraries, museums and collections of scientific data. Thus, archival institutions can collaborate with organizations in these other domains to develop from the enabling technologies solutions for long-term preservation and access.   In the archival domain, the promise of the persistent object preservation method has been demonstrated in several empirical tests, applying the method to a variety of collections across a broad quantitative scale.   These demonstrations involved bringing the collections into the archival information system from external sources; examining the documents, databases, images, geographic information systems and other digital objects tested in order to generate XML models; transforming the records and capturing collection organization according to these models; storing the transformed collections and related meta-data; and retrieving and presenting the preserved records using technologies completely different that those which had originally been used to create and store the records.

Conclusion

The persistent object preservation method offers several advantages to archives.   It provides a coherent and comprehensive framework that can be specifically tailored to archival requirements.

---

[29]Thomas A Phelps and Robert Walinsky. Multivalent documents: anytime, anywhere, any type, every way user-improvable digital document system. <Http://elib.cs.berkeley.edu/>

[30]Reagan Moore, Chaitanya Baru, et al. Collection-based persistent digital archives. D-Lib Magazine. March and April 2000, vol. 6, nos. 3-4. <http://www.dlib.org/dlib/march00/moore/03moore-pt1.html.> and <http://www.dlib.org/dlib/april00/moore/04moore-pt2.html.>

Through abstraction of the context, structure and appearance of the contents of digital objects, it provides a single, but highly adaptable method that serves at once the need for preserving authentic electronic records over time, for adhering to archival principles, such as provenance, and for performing core archival functions. Moreover, the persistent object framework permits the simultaneous adoption of other techniques if the need arises. Clearly a substantial amount of research, analysis, testing and evaluation needs to be completed before this method reaches its full potential. Nonetheless, the positioning of this method in the center of major developments in computer science and information technology offers great potential for making of electronic records not so much a problem for preservation, but an opportunity for archives to achieve their objectives to a greater extent and at a higher level than has been possible before now.