# Preservation of a Natural Electronic Archive

*Maria Esteva; School of Information, University of Texas at Austin; Austin, Texas, USA*

## Abstract

*This paper presents a case of analysis and preservation of a digital archive. The attributes characterizing the archive at hand led to the development of the concept of natural electronic archives that would allow transforming the archive into a unit of analysis. The results of the analysis directed the design of a preservation strategy aimed at maintaining evidence of the archive's formation process and documentation of the technologies involved.*

## Case Study

The digital archive presented here belongs to a philanthropic foundation whose activities in support of the arts and cultural patrimony, sciences and education, and social promotion in Argentina span from 1985 until its closing in December of 2005. During 20 years, the foundation invested net 98 million dollars in grant programs and own initiative projects and elicited investments doubling that amount in matching funds [1]. Founded by business men, the organization included academics and experts in the interest areas in its staff and board to shape and audit its programs. Through the years its personnel remained relatively stable, and never exceeded the number of 25 full time staff members. Programs were carefully planned and budgeted, and the staff worked efficiently to accomplish them with the least overhead and in timely fashion. Anticipating the end of the foundation, I approached its authorities in 2003 with the proposal to prepare the institution's records for their legal retention and eventually their long-term preservation.

Through 2004 and 2005 I appraised the paper and electronic records and coordinated the inventory, preservation housing, and off-site storage of the paper portion. For the last two years I have been involved in planning and supervising the preservation of the electronic records. The appraisal process showed the complexities and limitations of private institutions when they are simultaneously faced with the task of closing down and deciding the fate not merely of their archives but perhaps of their place in history. In addition to these particular issues, this archive shows general aspects of others formed during the late 1980's and 90's, a period during which as information technologies were being massively adopted in the work-place, problems were compounded by the nature and conditions of its electronic records.

Records creation and record keeping decisions and practices in the organization never involved an archivist or a records manager. Originally in tune with secretarial work-practices of electric typing machines, paper workflow registries, and centralized filing systems, the incorporation of networked computers, databases, and shared directories introduced new work dynamics and modified the original structure of the paper archive. Along the way, transitions from paper to digital transactions were decided with fearlessness and unknowns, and problems were solved on a first come first serve basis.

In the late 1980's the organization installed a networked server with PC work-stations to maintain shared applications and its grant tracking and financial database systems. In the early 90's a shared directory was implemented on the server to improve collaborative work. In it, each staff member had a virtual folder to store the files that they generated and received, the majority of which belongs to the Microsoft Office family and a minority to lossy compressed image formats. Within each folder, individual recordkeeping practices ruled how records were named, used, organized, and what was kept or discarded. With the exception of some restricted folders, staff members could access almost everybody else's files.

The same idiosyncratic style ruled IT operations and the way in which records and applications were managed in the networked server. IT guidelines and documentation were scarce, the latter reduced to current hardware inventories and software licenses. In the server, directories labeling was cryptic and granular details related with security policies and maintenance procedures were in the head of the systems administrators. Over time and after their use was discontinued, applications were not always removed from the server. No user manuals for the custom-made databases were produced and instructions were passed along from one employee to the other.

After the first 9 years during which DOS was used as the common platform for applications in the server and in the PC work-stations, the organization moved in 1996 to a Windows environment. At that point, the data from the databases was migrated to a newly designed Windows compatible system. DOS compatible text records on the other hand were transferred to the new environment but not migrated to Word for Windows versions. Other DOS applications and tools –including the old database system– were also transferred to the new server and remained there un-used. Around 2001, as part of its closing process, the institution discontinued major computing upgrades. At the time of cessation, the IT infrastructure consisted of a Windows NT 4.0 server, a combination of 1998 and 2000 operating systems in the work-stations, and Microsoft Office 97 applications.

After closure, for accountability and legal reasons, the institution's records have to be retained under restricted access for the next 10 years. While the paper portion is stored in a commercial records management facility, it was decided that the electronic one would remain under the custody of the last president of the institution. In this juncture, the goal of the digital preservation strategy was to establish and maintain the authenticity and integrity of the records and databases while guaranteeing that they remain accessible. Also, it had to account for limited financial and computing support during the next 10 years. To devise such strategy, a thorough analysis of the contents of the server and their dependencies was undertaken.

The analysis, had to overcome documentation gaps and identify what within the server contents constituted the archive.

## Discovering the archive

The approach used to study this case is partly taken from reflexive archeology. The goal of reflexive archeology is to contextualize the research methods. This is, to study a site –in this case the contents of a server– from the perspectives of the different contexts involved. In this way, it is expected that the conclusions that are derived will respond to all of them [2]. In practice, using reflexive methods meant first, gathering information from the technical and social contexts involved in the creation and use of the electronic records and systems in the institution. Continuing with data analysis, it was necessary to go back and forth between these contexts to determine relationships and points of conflict and convergence.

To gather social context data I interviewed 16 former staff members and 4 IT consultants and systems administrators that worked for the institution in various periods. The number of staff members interviewed represents 70% of those whose records are in the networked server. During the interviews I learned about the different record-creation and record-keeping practices, how subsequent information technologies were used, and who had been involved in selecting and implementing them. The interviews provided insight about work-practices and the relationships between organizational functions, staff members, use of the databases, and types of records produced. However, many interviewees did not know much about the information technologies used over the last 20 years, or could not remember the names, dates, and versions of hardware and software with precision.

To learn about the technological context I conducted a systematic analysis of the networked server. Using an inventory form I recorded upper level directories' names and provenance, encompassing file dates, file formats, record types, existing applications, and if it was possible to render the files or run the applications. In many cases, the contents of the directories were easily identified, in others, I had to look up information about file extensions in different sources or ask help from former systems administrators.

Through automatic metadata extraction from a sample of 700 files from the shared directory I could determine the name and version of most text editing software used in the institution from 1991 to 2005. DROID was the tool used for this purpose [3]. For the cases in which the results were tentative –files created with DOS compatible software– I used FileMerlin™ conversion tool to confirm precise software version [4]. Lastly, the accounting books in which major computing purchases were recorded with some level of detail, constituted an accurate source to learn about dates in which equipment upgrades were made.

As I was conducting my observations, I started to think of the networked server as a work-place environment which, by the end of the institution's existence, resulted in a digital archeological site. Over the years, all the staff members in different capacities and making different decisions had a part in its generation. Without much afterthought and in a natural way, records and applications were dropped, kept, and deleted, resulting in strata of different types of digital cultural material.

The data was recorded and organized in different documents to facilitate analysis: a staff timeline includes names, roles, and work-periods of each staff member whose records are stored in the networked server; an interview sheet contains staff members' recollections and perspectives on records-creation, record-keeping and use of information technologies; a technical metadata timeline contains hardware and software names and versions used in the institution during its existence. In the latter, I included the dates of commercial release and product discontinuation, and dates of institutional implementation and cease of use in order to perform cross-referenced analysis of the broader context in which technology decisions were made.

These different sources complement, confirm, challenge, and inform each other and the contexts that they represent. Compilation of the list of the technologies used in the institution was only possible by merging information obtained from the different sources. For example, Dbase query scripts found in the server confirmed interview accounts regarding the purposes served by the early DOS compatible database and clarified the encompassing dates in which the system was implemented. While many non-functional, the existence of old applications informed about what was used to manage files in the server and the staff schedules, and about the progression of accounting software used in the institution. For future exploration, the way in which the staff organized and named their files in the shared directory will allow uncovering as many recordkeeping patterns –or non-patterns– as amount of members in the organization. Finally, the different products of analysis merge in a narrative that combines the technological and the social contexts involved and describes the nature of the phenomena at hand.

## A natural electronic archive

I concluded that the manner in which records and tools had been kept in this server could not easily be ascribed to digital archiving models currently discussed in the literature. These focus more on the creation of sound electronic records, the design of electronic record-keeping systems, and on institutional repository archiving models than on the way in which digital archives are actually created [5],[6], [7], [8], [9]. The natural archive concept builds partly on that of "natural collections" proposed by Phillip Cronenwett to describe collections of literary manuscripts that are not fragmented as they leave the hands of their creator [10].

Creation of "natural electronic archives" involves a set of *ad-hoc* practices developed as people adjust to and learn how to use information technologies. A natural archive is not designed or managed by records managers or archivists. Instead, it is what those working in institutions, in different capacities, using different technologies, and making decisions, make of it. In a natural archive each record creator decides on naming conventions and organization for files and folders, spontaneously or consistently, according to individual mnemonic rules or the spur of the moment. Within the directories and sub-directories, images, spreadsheets, texts, web pages, databases, back-ups, email, scripts and applications live together in organized or

disorganized fashion, sometimes misplaced, and without descriptive clues.

Record creators create, re-invent, and leave behind natural archives or parts of them in iterative fashion. Within these iterations, archives evolve towards more structured forms or become interrupted. During these processes, pieces of the archive are left like discarded artifacts in an archaeological site because there is no time or need to go back to old files that cannot be found or applications that cannot be opened. This constant halt and advance is intensified by emergent technologies and ways of creating and storing records, and by the appearance of new users. At any time, a formerly new way of constructing the archive may be superseded; its criteria, passwords, naming conventions, and logic left behind. A natural archive is an aggregate –or better– an accretion of trials and errors. The evidence of this resides in the vestiges of directories, files and application left in the storage space and in the lack of consistency in naming, organization, versioning, keeping, and discarding that characterizes them.

Consequently, records within a natural archive are difficult to identify and lack formal documentation. This creates doubts about their authenticity and their capacity to provide evidence, and challenges traditional notions about the way in which they have to be preserved.

## Minimalist Preservation Strategy

As a response to what was observed during the analysis, a preservation strategy that focuses on maintaining the natural archive's formation process was devised. Formation process is an archaeological concept applied to the study of archival and museum collections. It comprises the combination of people, tools, and circumstances – environmental or social – that affect a collection or a system upon its creation and over time. As explained by Schiffer, "In order for material entities – documents, photographs, the ordinary and extraordinary things of everyday life, even people– to serve as evidence, they must persist over time and be found and studied by investigators. Formation processes create the pathways leading from past behaviors to evidence of them in the present." [11]

The basic premises of the preservation strategy are:

- Maintaining evidence of the way in which the archive was formed.
- Maintaining the integrity and authenticity of the records over time.
- Maintaining access to records and systems at a minimum during the required retention period.
- Continuing documentation of the evolution of the archive.

A dark archive was created on a new server hardware with RAID 5 configuration to host the contents of the old server in the same directory structure in which they existed. The server's environment was prepared by installing the current Windows 2003 server system and current versions of compatible applications and file viewers to render most files present in the collection. Windows was the selected platform because it is the system in which the databases and applications function. The purpose of purchasing new hardware was to increase data security and prevent equipment failures in the next few years. The reason for installing up-dated software was to catch up with the new hardware and avoid the shortcomings of maintaining obsolete technology [12].

Foreseeing limited use, to increase security and to avoid maintenance overhead, the archival server will be kept off-line and un-plugged [13].It will be turned on only by authorized users to retrieve information when and if needed, and to perform regular server maintenance routines and records' integrity check-ups. An audit and control software, was installed to perform file integrity checks through hashes, detect changes in files, and monitor access to the server [14]. The old server with its contents will also be kept "as is" for the purposes of studying technology obsolescence until mechanical failure occurs.

Before transfer, rendering of text and image files, and functionality of the databases were tested to determine whether migration to the new server system would affect them in any way. The grant tracking database created in Clarion for Windows 2.2 in 1996 functioned without problems in the up-dated environment [15]. Test results for Microsoft Word documents indicated that rendering of files from 1992, 1993, and 1994 created with Microsoft Word versions 5.0 and 5.5 for DOS, does not change when opened with Microsoft Office Word 97 for Windows –installed in the institution's work-stations –nor with Microsoft Office Word 2003– installed in the dark archive. In both cases file rendering is equally defective due to differences in the underlying encoding of Spanish characters and formatting present in the original Word for DOS files. Rendering of the rest of the Microsoft Word files created since 1995 in different versions for Windows do not present problems. These results indicate that the way in which records were available to the users has not changed since the institution moved to a Windows environment eleven years ago. At that point it was decided that migration of files created before 1995 would not be made. In the future, the documentation produced will help determine migration priorities.

A protocol was written specifying operations to insure maintenance of the original file properties during transfer, monitoring of the content's integrity over time, and documentation of every activity to be performed on the server. Transfer procedures include: automatic inventorying of the contents of the old server (pre-transfer) and those of the archival server (post-transfer), virus check, and an initial hash generation for all the contents in the new server. In December of 2006, transfer of the contents was carried out over the local network in the same directory structure as they existed. The documentation stating the steps involved in the transfer and the inventories describing the archive, was given to the archive's custodian and saved in a specific directory in the dark archive.

The dark archive's maintenance routine will entail annual monitoring of the integrity of the contents through the reports and logs generated by the server auditing software. Post-transfer hashes will be compared with new ones in order to determine if file modification has occurred in which case the server access and file change logs will pinpoint who did it and when. Copies of the archival server contents were made onto DVD and recommendations where made to store them off-site. In the case of hardware failure or file changes, the system will be restored from these back-ups. One of the copies will be used for annual verification of the DVD media to decide if it needs to be

refreshed. Finally, a directory structure and a file naming scheme were created to include the monitoring documentation produced over the years.

As new hardware and software become available, future up-grades will be evaluated in relation to file migration needs, and to final decisions about the destiny of the archive beyond the 10 year retention period. Currently, the content of one of the DVD back-ups was transferred to a secure server available for my research. A sample of records and the databases available in this copy will be used for file rendering and functionality tests in order to determine migration needs. The whole strategy is underlined by bitstream preservation. In addition to future migration decisions, the natural archive as it was transferred to the dark archive will always be kept in its original form.

## Conclusions

Considering the server akin to an archeological site allowed evaluating the technological dependencies and social uses of records and systems in ways that might not have been possible if applications, records, and systems found "in the site" were dealt with separately. As a result, the concept of a "natural archive" emerged. I suggest that this concept applies both to the case study at hand, as well as to archives of public or private persons and institutions showing similar characteristics.

The exploration suggested a preservation strategy that aims to maintain the context in which the records were created and used and consequently the evidence of the archive's formation process. It stresses an audited transfer to a new and compatible server environment configured to protect the authenticity and integrity of the contents, and involves minimal maintenance and annual monitoring to assure continuing access to the records.

## Acknowledgements

Thanks to Dr. Patricia Galloway, Dr. Victoria Horwitz, and Aaron Choate for reviewing the manuscript.

## References

[1] Information about the institution was obtained from the foundation's final report issued in 2005. Full reference is not disclosed due to confidentiality agreements.

[2] Ian Hodder Ed., Towards reflexive method in archaeology: the example at Catalhoyuk (BIAA Monograph No. 28, 2000)

[3] The National Archives, DROID: Digital Record Object Identification ,Version 2.0, (Computer software, 2006) http://droid.sourceforge.net/wiki/index.php/Introduction

[4] Advanced Computing Innovations, Inc., FileMerlin, Advanced File Conversion Software (Computer software, 2006) http://www.acii.com/fmn.htm

[5] Authenticity Task Force, Requirements for assessing and maintaining the authenticity of electronic records (INTERPARES, 2002) http://www.InterPARES.org/_file.cfm?doc=ip1_authenticity_requirements.pdf

[6] D. Bearman, & J. Trant, "Electronic records research working meeting May 28-30, 1997: A report from the archives community" D-Lib http://www.dlib.org/dlib/july97/07bearman.html (1997, July/August).

[7] R. J. Cox, "Electronic systems and records management in the information age: An introduction" ASIS, 23(5). http://www.asis.org/Bulletin/Jun-97/cox.html (1997, June/July)

[8] L. Duranti, T. Eastwood, & H. McNeil, Preservation of the integrity of electronic records. (Boston: Kluwer Academic , 2002)

[9] R. Jones, T. Andrew, & J. MacColl , The Institutional Repository ( Oxford, UK: Chandos Publishing (Oxford) Limited, 2006)

[10] P. L. Cronenwett , "Appraisal of literary manuscripts" In Nancy E. Peace, Archival choices: Managing the historical record in an age of abundance (Lexington, MA: Lexington Books, 1984) Chap. 5, 105-116

[11] M. B. Schiffer, "Formation Processes of the Historical and Archaeological Records" In David Kingery Ed. Learning from Things: Method and Theory of Material Culture Studies (Smithsonian Institution Press: Washington and London, 1996) p. 74

[12] Cornell University Library, "Digital Preservation Strategies", Digital Preservation Management: Implementing short term strategies for long-term problems, http://www.library.cornell.edu/iris/tutorial/dpm/terminology/strategies.html (2003).

[13] New data storage technologies offer systems that get turned off and on according to needs to avoid wear and save energy. See Copan Systems, MAID Technology (Computer software, 2007) ,http://www.copansys.com/architecture/index.shtml

[14] Tripwire, Tripwire for Servers/Manager, (Computer software, 2006) http://www.tripwire.com/products/servers/index.cfm

[15] Clarion is a database development technology already in version 6. See SoftVelocity, Clarion, (Computer software, 2007) http://www.softvelocity.com/clarion/c6.htm