Considerations on the Preservation of Base Digital Data of Cultural Resources

Nicola Barbuti

Dipartimento di Scienze dell'antichità e del tardoantico, University of Bari, Italy
nicola.barbuti@uniba.it

1 Introduction

This paper does not aim to thoroughly list and discuss issues which are well-known in research circles working towards defining the correct strategies for the preservation of cultural heritage digital resources, nor is it an attempt to suggest possible solutions to the current problems, as this is a burdensome task which has been tackled by individuals much more professionally and theoretically qualified than myself.

This paper does, however, provide food for thought and raises pertinent questions which come up on a daily basis for those who through their research or work encounter the aforementioned issues. The sole aim, if at all possible, is look upon the issue from a different perspective to those adopted in the discussions and theoretical debates on the delicate and still unresolved question of digital preservation.

2 State of the Art

The common definition of digital preservation is 'the collection of activities and instruments which guarantee that digital documents are kept accessible, usable (legible and intelligible) and authentic (unambiguously identifiable and intact) in the medium and the long term, in a technological environment which is definitely different from its original environment.'[1]

For years the pressing need for planning common and definitive strategies has been the source of major concern and repeated appeals by archivists and librarians the world throughout. Lately, with the recent heady progress in the adoption of digitisation even in institutional and public administration settings, this has become a primary need and a real emergency. There is in fact a growing awareness that if we continue with the current intuitional and scientific confusion and indifference, the legacy of knowledge we will leave future generations regarding the beginning of this millennium will be next to nothing.

It is worth recalling the extent and the significance of the problem within the scientific community with a summary of some of the most important theoretical contributions published on the subject in recent years.

[1] «L'insieme delle attività e degli strumenti che assicurano che i documenti informatici siano mantenuti accessibili, utilizzabili (leggibili e intelligibili) e autentici (univocamente identificabili e integri) nel medio e nel lungo periodo, in un ambiente tecnologico certamente diverso da quello originario.» M Guercio, *La conservazione digitale nello scenario europeo e internazionale. Principi, metodi, progetti*, Rome, November 2003, p. 2 (url: http://eprints.erpanet.org/archive/00000064/01/e-book.pdf).

M. Agosti et al. (Eds.): IRCDL 2012, CCIS 354, pp. 13–20, 2013.

It is readily apparent that the studies, proposals and speculations put forward by renowned scholars and researchers such as Mariella Guercio, Perluigi Feliciati, Stefano Allegrezza and Paolo Franzese, just to mention a few, are all bound by a common thread: they note with clear and incontrovertible argumentation and documentation that in contrast to the enormous flurry and mobility at the international level, the problem of preservation in Italy is still unresolved because it is ignored in its entire extent. The debate and fragmentary projects which are underway are suffering under the conditions of complete backwardness due to the incapacity and, worse still, the unwillingness of the political institutions and universities to take on the task[2]. Not to mention the perennial shortage of financial resources which are made available for studies and projects in the sector.

Taking it for granted the we all consider the proposals and the considerations outlined in the abovementioned studies of vital interest, I would now like to briefly outline some considerations made on the topic from 2008 onwards which emerged following the setting up of a project for the creation of a multimedia library specialised in ICT research for archival and library cultural resources[3] and the subsequent setting up of a spin-off activity which I currently head in the ICT sector for cultural heritage, with specific reference to archival and library cultural resources.

3 Questions without Answers

As I stated earlier, the main problem which seems to plague digital preservation in Italy is that political and university circles in this country are not interested in committing human let alone financial resources towards serious cooperation in international initiatives promoting digital preservation, or when they do show an interest money is wasted without producing results worthy of mention, or the sum production is a total failure[4].

---

[2] In particular, see the following: M. Guercio, Archivi digitali e conservazione a lungo ter-mine. Un quadro di sintesi sulle strategie internazionali e nazionali, Archiexpò, 12-15 December 2006 (url: `http://ebookbrowse.com/archiexpo-guercio-ppt-d26376049`); P. Franzese, Archiviazione e conservazione del delle risorse digitali. Les archives électroniques. Manuel pratique edited by the Directorate of the Archives of France (February 2002), Rome, October 2006 (url: `http://media.regesta.com/dm_0/ANAI/.../000/.../ANAI.000.0113.0012.pdf`); S. Allegrezza, Informatica di base. Conoscere e comprendere le risorse digitali nella società dell'informazione, Edizioni SIMPLE, Macerata, 2009; P. Feliciati, Il nuovo teatro della memoria. Informatica e beni culturali in Italia, tra strumentalità e sinergie, «Il Capitale culturale. Studies on the Value of Cultural Heritage», vol. 1, 2010, p. 83-104.

[3] The "Unknown Heritage" project, together with the Unknown Heritage workshop of which I am currently the Chief Scientist.

[4] P. Feliciati, p. 86: 'This instrumental relationship [between information processing techniques and activities related to cultural heritage], due to haste, limited competence or lack of vision which goes beyond the use of extensive resources to obtain – at the best of times – special effects which last no more than the length of a legislative period, has to date only widened the gap rather than produce synergies'. He adds shortly thereafter (p. 88): 'Except for some sporadic examples of excellence, the rather disapproved "cultural deposits" are noteworthy for the waste of resources in projects which are ad hoc, isolated and without any worthwhile duration or real utility regarding the digital objects obtained … Moreover, they are all projects which are cited in the literature as poor examples of long-term preservation of digital resources, since with only a few exceptions most of the data which were gathered at such a high price are by now totally lost or useless".

Another reason which has been put forward is the lack of professionally trained personnel able to take up the arduous challenge, due to severe inadequacies in the national university system. Still today these inadequacies have not been resolved[5].

Nonetheless, while recognising the clear distinction currently present between digital humanities in Italy and abroad, it can be said that there has been increasing involvement at the international level of archiving and library circles in studies and research on information science and digital libraries, all the while noting the lamentable backwardness characterising the Italian scene with respect to international scenarios, even though in recent years a more solid foundation seems to have been laid down[6].

On the backdrop of this scenario, when the pilot project "Unknown Heritage" was launched in 2008 the first problem we faced regarded the definition of criteria able to guarantee an acceptable percentage of probability that the databases planned as the output of the project would be usable on the web for at least a five-year period after their creation.

After several months spent evaluating the state of research regarding the market sector particularly in Italy, we were forced to accept the reality which had been so well described in the studies mentioned above: approximation and fragmentation of the scientific research, inhomogeneity in the best practices theorised even at the international level, products and services touted as innovative and revolutionary which after closer examination proved to be totally incapable of satisfying the needs of university project, and so on.

In short, the problem was still there and remains unresolved still today. Nonetheless, I continue to surf the web on a daily basis, consulting specific bibliographies in search of signs foreshadowing the decisive momentum needed to achieve possible solutions.

At any rate, despite our perplexity and awareness of the risks we were up against, the project saw the creation of two databases of equal content but different software architecture which are both usable today on the web without them needing to be updated[7].

The problem of preservation has however become a pressing issue for me, in that as stated above the project gave rise to a university spin-off D.A.BI.MUS. s.r.l. – Digitalizzazione di Archivi BIblioteche e MUSei (Digitisation of Library and Museum Archives), the activities of which are digital ICT for cultural resources, with specific reference to archival, library and museum heritage.

Between 2010 and 2011, the spin-off in fact planned and created an innovative application which is currently pending patent. The application is a digital recognition

---

[5] P. Feliciati, p. 89, 90.

[6] P. Feliciati, p. 86, 90.

[7] On the pilot project "Unknown Heritage", see N. Barbuti, Valorizzare tutelando. Il Laboratorio multimediale e la banca dati digitale "Patrimoni Sconosciuti" dell'Università di Bari, «Biblioteche Oggi», No. 3, April 2011, pp. 38-44. Four years down the track the two databases are still visible and totally accessible at the following URLs:
http://digilibrary.patrimonisconosciuti.uniba.it e
http://virtualibrary.patrimonisconosciuti.uniba.it

suite, which includes functions of Intelligent Character Recognition, Intelligent Word Recognition, Optical Character Recognition and Graphic Pattern. The application is capable of operating with a high level of efficacy and efficiency on the basis of digital data of antique printed documents, manuscripts and books[8].

The main difficulties which arose during the planning phase of the system and which were made apparent by various parties during the presentation of the suite include the question of its spatiotemporal duration and the need for multiplatform functioning. We are currently working towards achieving these two objectives: the functions of the suite, which currently are only available for Windows, will be extended for Unix/Linux and MacIntosh platforms, and the algorithm will be developed so that it functions on all image formats currently in use, but above all on the Flexible Image Transport System (FITS) format.

The choice of insisting on this image format as a standard for the development of both the university research and the products of the spin-off is well grounded. FITS has in fact been in use at NASA for almost 50 years as an image format for astronomical photographs. Indeed, it presents all of the necessary characteristics to guarantee the base digital data will enjoy the maximum duration and usability over time without too many risks of destruction and without excessive costs for maintenance or updating.

FITS is a non-proprietary image format which we believe, thanks to its characteristics of usability and portability in time and space, could validly constitute a real starting point for developing strategies aimed at definitively resolving the problem of the preservation of digital originals and the metadata associated with them, and therefore their safety and consultability over time.

Some might argue that it is uncertain whether this image format, which is as I mentioned commonly used for astronomical images, is similarly valid for the reproduction of other materials and in particular cultural resources. We counter that the Vatican Library, after years of tests and checks, was the first in the world to adopt FITS as its main format for the project of digital reproduction of its immense manuscript collection. Based on this initiative, on 5 July 2012, during the EWASS 2012 Conference held at the Pontificia Università Lateranense, the Vatican Library organised a special session entitled *Long-term preservation... from the stars? File format assessment and technical issues in preservation projects for cultural resources*, in which the excellent results were presented. Indeed the Vatican Library has rightly become a standard bearer with its adoption of FITS as an image format shared at the ecumenical level.

Of course, it remains to be seen whether FITS is a format which can also be used for the production of native digital document. Nonetheless, the initiative of the Vatican Library marks a watershed and a significant step forward in the definition of the strategies for digital preservation, and it is worth undertaking serious research to ensure that this beginning does not remain, as often happens, an isolated phenomenon.

---

[8] On the function of graphic pattern, implemented by the spin-off in the setting of a project currently underway in cooperation with the Vatican Library, the following paper is currently undergoing revision: N. Barbuti, T. Caldarola, Graphic Matching in Historical Manuscripts.

It is worth examining at this point the concept of original mentioned above in that, when tackling the problem according to an archival/library science approach, the question is not so much one of the preservation of the image as an end in itself, but rather the preservation of the originals, which is a pivotal, inescapable and irreplaceable concept in the doctrine mentioned above.

Indeed, the distinction between original and copy, which is well defined for analogic objects, is extraordinarily subtle for digital resources. From the moment of its creation an electronic or digital document undergoes rapid modifications which irreversibly alter its original structure. From the moment of its creation to the moment of its publication the so-called 'definitive' document ends up being the copy of various and subsequent copies germinating from the original, which has been inevitably and irreversibly lost along with its subsequently revised versions. The document which reaches the end user is only the final interface, and nothing is left of its construction and the original which preceded it in the first phase of its creation. Nonetheless, the definitive document is by definition considered to be the original, and as such it is archived on a stable support and preserved so that consultation and manipulation do not compromise its integrity and possibility of survival over time.

We are all well aware that there are various causes of irreversible destruction of an electronic or digital document/archive which have a decidedly frequent incidence[9].

To avoid the problems due to the different forms of obsolescence which could compromise their survival over time, the base data need to undergo a process of periodic updating[10].

Now, reflecting for a moment on this necessary procedure, the following question comes to mind: every time that a digital document undergoes updating, what happens to the original? Is the document created by the update considered the original, or is it a copy of the previous document which nonetheless maintains the security data and the metadata unchanged? And above all, are the metadata and the security data truly preserved in their entirety and perfectly identical in the new document? Several doubts remain, and we are patiently awaiting answers which are clear, thorough and above all definitive.

In order to clarify what we believe to be the extent of the problem, let us now briefly examine the change in the transposition of the collective memory of human endeavour onto a transferable support, and how this change has today reached a point of no return, which is slipping through our fingers unnoticed, or at least unmentioned.

---

[9] We are referring to the three major problems which create difficulties for the preservation of digital content: obsolescence of hardware and software technology, obsolescence of the support and obsolescence of the formats. In addition there can be accidental causes, such as prolonged exposure to heat, or those due to inadequate environmental preservation. See in this respect Allegrezza, p.2.

[10] See S. Allegrezza, p. 4: 'Over the last fifteen-twenty years the problems of digital preservation have been tackled from numerous viewpoints and a variety of preservation strategies have been put forward suggesting a solution. The main ones include: output to analogue media; technology preservation; emulation; refreshing and migration; and digital archaeology'.

For thousands of years mankind has been an avid inventor of new methods with which it has entrusted its evolutionary history so that it can be passed on to future generations. And for thousands of years these methods have had the common characteristic of tremendous stability over time and space, characterised as they are by a rate of decay which is either virtually non-existent (stone, pliable materials) or extremely slow (papyrus, parchment, paper). All of these materials have proven their ability to resist natural calamities and the disasters of warfare and thus allow us still today to study their content and acquire even greater knowledge about who we are and where we come from.

With the advent of the analogic age between the 19th and the 20th centuries (wax, charcoal, vinyl, photographic film, audio and video), the ability of the materials to last over time has noticeably decreased with respect to the past, all the while maintaining still acceptable levels.

However, it has been the advent of the electronic and then the digital age, characterised by recording processes which are unstable and volatile by definition, which has marked the beginning of an irreversible disappearance of a significant quantity of contemporary collective memory.

Consider for example that while I am writing this paper, millions of billions of data created by mankind throughout the world are disappearing into the ether, taking with them the collective memory that they contain. Consider for example that this very paper has already been rewritten numerous times in many of its parts, its many errors and typos corrected in real time, such that in the end it has become a final and very different copy of what would have been the original project, if instead I had decided to first write it down with a pen and paper, and then in a word-processed document, and lastly made a comparison of the two products.

Electronic mail and on-line communication, chat and social networks have definitively blown away all practice of handwritten interpersonal communication, a heritage which for centuries has allowed us to understand the life and culture of those who with their lives and their works have written our history.

With the word processed document it is impossible to identify the path running from the gestation of any thought through to its birth, elaboration and publication. The digitisation of the public administration will make it impossible in a short space of time to reconstruct contemporary social, economic, health and demographic history. For future generations living in an age of hypertechnology even more volatile than our own, if that is at all possible, such a reconstruction could be useful for understanding their role on this planet.

This is not a post-apocalyptic scenario from some science fiction screenplay. Instead it is the reality which passes daily before our eyes and which for years we have called the technological miracle, without realising that we are passively succumbing to the rapid destruction of contemporary collective memory, a drama in which we are both actors and directors. If we fail to shore up this flood towards oblivion, we run the risk of becoming the historical age without a clearly identifiable past and with no collective memory of the present capable of creating a future: an endless present, the first and only true dark age in the history of humanity. The dark centuries of the Middle Ages will pale in comparison.

Therefore, we feel that the problem of digital preservation is no longer a problem of the capacity of real or virtual hardware space made available for preserving our digital collective memory. Nor is it a problem of distinguishing between an original and a copy, which nonetheless is a question of primary importance in the choice of what needs to be preserved what can be eliminated. It is not even a problem of the obsolescence of magnetic, optical or who knows what other type of support which will appear on the scene in the next five years or so.

The problem of preservation, a problem I have been wrestling with since I became interested in information technology for cultural heritage some ten years ago, is becoming something much larger. Indeed the problem involves society as a whole, in a setting where it appears there is ignorance of the fact that already a large part of what was created in the last 20 years was destroyed at the moment of its creation, and much of what has survived is being destroyed at a tremendous rate. It is a problem of preservation of collective memory in a digital format, and not one of the "simple" – but as we already know very complex – digital preservation. If we fail to carefully focus on this analytical perspective of the problem, we feel that it will be very difficult to develop common strategies capable first of restricting the flow and then of stemming it entirely with appropriate and timely planning.

Moreover, perhaps even as a result of the difficulty of framing the issue in its true extent, in this scenario researchers, scholars and operators active in the various information, cultural and administrative sector, who despite their daily ringing of alarm bells regarding the worrying situation and who apply themselves with a passion and in some cases a high degree of professionalism to provide hypothetical solutions to the problem, in reality seem more interested in justifying their own membership to their respective sectors and the supremacy of each over the others.

By now the scientific papers and studies and the research published on the different problems surrounding digital culture and its preservation are on the daily agenda, published above all by computer scientists and archivists/librarians who have chosen to broaden their knowledge of the new technologies and who have, so to say, lent themselves to computer science. However, an analysis of these studies reveals that the approaches to these problems and the possible solutions are still diametrically opposed. The computer scientists are sunk in their endeavours to develop algorithmic structures or theoretically perfect networks, which nonetheless are often practically unusable and therefore destined to remain contemporary pipe dreams which will soon be forgotten or overtaken by new theories, which will also be perfect and unusable. The cultural scholars are instead hunkered down in their own defensive positions built on the few and certain results they have obtained and their supposed eternal validity, voluntarily unaware that without programmes and projects built on synergies between qualified professionals originating from both sectors, those results will be destined to a life much shorter than what is need as they are the result of fragmentary, inhomogeneous policies often brought about by more of a need to put on a show than to really make cultural resources available to the present and future generations.

The approach to the problems of the preservation of collective memory in a digital format can only be interdisciplinary and must cut across the various cultural, scientific and social forces, and only with a major effort and results which are worthy of note

will it be possible to achieve a clear vision of what is happening, and as a result, provide the planning of preservation policies which are shared, certain, effective, efficient and long lasting in time and space.

In conclusion, let us finish with this bitter sweet literary digression:

> 'From the right the sound of a trumpet is heard,
> from the left [still no] sound is heard in reply.'[11]

And yet, faithfully we shall wait, and in the meantime we will continue do research, to study, to compare and to grow.

[11] Translator's note: Alessandro Manzoni, Il Conte di Carmagnola, «S'ode a destra uno squillo di tromba / a sinistra [ancora non, nda] risponde uno squillo», with the text 'ancora non' inserted in square brackets by the author. The verse refers to two armies facing each other on the battlefield. They appear to mirror each other, and it is noted that neither is an invading force – the question of brotherhood is raised as the armies are composed of Venetians and Milanese.