Digital Archiving Tomorrow: A Foresight

Lukas Rosenthaler
Imaging and Media Lab University of Basel
Basel, Switzerland

Abstract

A large portion of the audio-visual heritage of our time is already stored on digital media or will be transferred in the near future onto such media. Unfortunately, current digital media are even less stable than traditional media and therefore periodic data migration every few years is mandatory. This approach is very cumbersome and risky. However novel concepts show a way out of this dead end situation. The underlying idea is that the digital data should not be passively migrated, but it should actively migrate itself. The idea is to develop future archiving concepts which are modeled on nature, i.e. on the living system. The genetic code (which is basically a 2-bit code) is the basic information for all biosystems, and it archives the information about species very successfully. Computer viruses and worms, despite being a severe nuisance, demonstrate that artificial live can be very successful and is very difficult to exterminate. Instead of having a maleficent payload, viruses and worms could be beneficial carriers of archived information and thus turned into positive "beneware".

Introduction

Long-term archiving of cultural assets always has been a very difficult task. Many objects of cultural interest have not been designed with longevity in mind. For example photography is inherently unstable and all photographic material will decay with time. While B/W-photographs have an life expectancy of about 100 to 150 years (until the first effects of decaying become visible), color photography is usually decaying much faster. Even modern photographic material which is chemically more stable will decay with time. The decay rate can vary very much and is dependent mainly on storage conditions but also on used material, processing methods, handling etc.

Since many objects in archives are unique artifacts which are irrecoverable in case of damage or total loss, every handling of these always poses great risks for damage or total loss. Often, the handling is also impractical and cumbersome. Therefore many museums and archives are involved in digitization projects. Unfortunately not all types of cultural objects are suitable for digitization, and digitization is only able to capture a fraction of the essence of an object. In general the digitized object has lost all "materiality" – it's an immaterial representation of part of the original object. In order to use photography again as a well suited example, a digitized photograph can capture only the visual content of the original photograph whereas most aspects of the material such as physical properties (thickness, surface properties, smell etc.) will generally be lost.

Most digitization projects have been launched or are conceived without long-term archival in mind. These projects are started for improving access to the assets through databases and internet. But as soon as the first digital data arrives, the question of archival arises: the digitization process is slow, expensive and cumbersome and therefore will usually not be repeated in foreseeable time. As a consequence, most institutions involved in large-scale digitization processes suddenly are facing the problem of long-term archival of digital data.

In order to complicate the situation, many cultural assets created today are "born digital", that is are of digital origin. Especially visual arts such as photography, motion picture, computer animations, video etc. using modern technology result in "originals" of digital nature. This fact will increase the pressure for archives to find solutions for digital long-term preservation "real soon now".

Digital Long-term Archival

From our daily experience, digital data seems to be very volatile and unstable. Everybody working with computers of any scale has had the bad experience of data loss, be it a word-processor document that becomes unreadable, an external storage medium that cannot be accessed anymore etc. It looks like "long-term archival" and "digital" are diametrically opposed concepts. However, the digital domain defined as the numerical representation of real world objects, offers some unique characteristics which make digital data especially suitable for long-term preservation. These characteristics are

Reproducibility

Because a digitized object is reduced to a list of numerical numbers, it can be copied - *cloned* - without any loss of information. In fact, for digital data, the notion of an *original* loses its meaning: there are no originals in the digital domain, only undistinguishable data files*.

Distribution

Digitized data is immaterial can be distributed without the effort to transport matter from point A to point B, and this with speed of light. Recent technologies such as BitTorrent[?] demonstrate that an efficient transfer of large amount of data over the Internet on a large scale is possible.

Media Independence

Digital data can be recorded virtually on any media. The 0's and 1's can be engraved into stone, etched in all kind of metals, recorded as magnetic marks on magnetic surfaces or be represented by single atoms on the surface of a gold using methods of nanotechnology[?]. The transfer in-between different media can be achieved with zero loss.

The combination of these core qualities make digital data especially suitable for long-term archival. Digital data can be stored with high redundancy, since copying with zero loss and distributing the data all over is possible. Copying the data from one storage medium generation to the next (*migration*) makes the unavoidable degradation of storage media irrelevant, provided the copying process is done as long as the deterioration still allows for a correct reading of the digital data. From a theoretical point of view, only digital data has the properties for a true long-term archival - all analogue data will deteriorate so much given enough time that it will become useless.

However, digital long-term archival still poses almost insurmountable challenges. The root of these problems lies in the fact that digital technologies are still in their infancy and immature compared to the established methods of archival. In addition, the rapid advance of the technology, while doubling the processor speed about every two years and increasing storage density exponentially (Moor's Law[?]), poses enormous compatibility problems in-between different generations of hardware and software. All types of storage media are prone to this problem. For example each new generation of magnetic tape of the DLT--family offers higher (often doubled) capacity than it's predecessor at the cost of a partial incompatibility. As a rule of thumb, the most recent reading/writing-equipment is able to read/write media one generation back, to read media two generations back, and cannot deal with older media types. This rapid change of technology is more limiting to the life-span of digital data than the actual aging of the physical recording medium.

As a result, the following fundamental methods for long-term archival in the digital domain are possible:

Preservation of the Hardware

Not only the storage media such as magnetic tapes, disks etc. are archived, but also the machinery and peripherals to read and manipulate the data has to be preserved in working condition. While this might work for older machinery, it is impossible for highly integrated electronic devices. If for example a chip of a tape controller is broken and there are no more replacement parts available, the whole machinery is useless and beyond any repair. Therefore we consider this approach unpracticable.

Emulation

The software and to some extent the hardware of obsolete computer system can be emulated on modern computers. In very limited cases, emulation can be a viable way for preservation. Especially software such as games etc. can be preserved by emulating the hardware and software of passed computer systems on new computers. This method however has some serious drawbacks:

- the emulation software has to keep up with the advance of computer technologies. This can be very difficult and costly. Otherwise, software emulators have to be used recursively (e.g. run Pac-Man on an Apple II emulator on a Mac-Plus emulator which itself runs on a Windows PC). Both ways unpractical and error-prone.
- Hardware peripherals are very difficult or even im−possible to emulate. For example, early computers used audio cassette tapes as storage medium. Even the best emulation on a modern PC is not able to read such old media.

Therefore we consider emulation generally not as a suitable method for long-term archival of digital data.

"Eternal" Media

The "eternal" media approach requires the digital data to be recorded onto the most robust and durable media available. We consider a recording medium "eternal" if it offers a life span that is approx. one magnitude larger than the life span of the original object. If we consider the life span of a normal digital medium to be about 5 to 10 years, an "eternal" medium would offer a longevity of 50 to 500 years. There are several attempts to create some sort of eternal medium, among others the Rosetta project of the Long Now Foundation[?].

The Imaging & Media Lab of the University of Basel very recently started a project named *PermaBit* using conventional B/W- or color-microfilm to store digital data. The basic idea is to expose the digital code as sort of a 2−dim or (in case of color-microfilm) 3−dim barcode using a laser-driven film recorder. Microfilm is known to have a durability of about 500 years and is a well known and understood medium. Combining a digital recording technique with the properties of microfilm will lead to a storage medium which can be considered "eternal" along the above definition. The recording method will take into account the special properties of microfilm in order to allow easy decoding[†]. Using redundancy, error correction codes and spacial distribution of the digital bits, even a locally damaged microfilm will allow full reconstruction of the digital data without loss. This approach has also the advantage, that analogue and textual information can be stored in a human readable form together with the digital data. For example an image might be represented as analogue "thumbnail"-image and as digital dataset on the same piece of microfilm, together with a human readable description of how to decode the digital data.

Migration

In the context of long-term archival, migration is defined as the process of reformatting (if necessary) and copying the digital data onto new media. Migration is a *periodic* task which has to be repeated before the media and formats in use get obsolete, and before the media show aging effects. For current magnetic and optical storage media, a migration period of less than 4 to 8 years seems to be necessary. The migration process has to be performed very carefully, because it is a zero-tolerance process. At the end, the content of the migrated data must be, bit by bit, identical to "original" data. If a reformatting is necessary, the new format as to be chosen very carefully in order to avoid any information loss during reformatting. Therefore migration is a very difficult and costly process which requires a lot of technical knowledge.

Current Developments

As we have shown, there is no way to evade digital long-term archival on the long run, but it poses still a lot of questions and technical and methodological challenges. We consider only "eternal" media and migration as viable methods which are worth further considerations. Preservation of hardware and emulation may work only in very limited and special cases. However, both "eternal" media and migration lack essential features to be used on a large scale, and both methods as in use today are too complicated and too expensive. Therefore our current research is focused onto these two archival methods.

PermaBit

As described above, PermaBit uses microfilm as medium to store digital data. Microfilm has a known stability, is relatively cheap and readily available. The read-back of the data must be possible with relatively simple equipment (e.g. a light microscope). To fulfill the highest needs of archives regarding to future readability, the code in the media domain will be designed as self-explaining as possible. Read-back instructions will be stored on the media as well as an error correction description. The decay of the photographic material (bleaching, dirt, dust, scratches) is an inevitable process and must be taken into account.

Distarnet

The basic concept of Distarnet has been presented at the IS&T archiving conference 2004[**?**]. Distarnet is being implemented[‡] as a distributed, self-organizing and optimized peer-to-peer (P2P) storage network with a high redundancy, based on an open protocol using XML technology. The data to be archived will be distributed in redundant copies among all participants parties. It will be assumed that the data will be transported over untrusted channels and stored on untrusted nodes. Therefore, strong cryptography may be used to guarantee the integrity and privacy of the archived digital assets. The fundamental idea behind Distarnet is that institutions with long-term archival needs of digital data collaborate in order to build a geographically distributed, Internet based archival system. It has been of primary concern in designing this system to minimizing the risk of data loss due to a catastrophic event (on a local or even regional scale such as a devastating earthquake) or false manipulation. In addition the archival network has to cope with the rapidly changing standards of information technology and preserve readability of the data beyond changes of software and hardware technology.

Distarnet therefore relieves the institutions from the burden of an explicit migration. Given that the Distarnet-protocol is implemented on the most recent hardware, an outdated storage node can be taken offline, and the new node is going online is automatically integrated into Distarnet and filled with data. As long as there is no reformatting (on the file-format level) necessary, migration becomes almost a non-issue.

Future Developments

All archiving methods for digital data currently available require special institutions taking care of the archival process. Even if methods like Distarnet facilitate the archiving process considerably, any short interruption in this process will lead to the unrecoverable loss of valuable data.

One migration too late, one year without funding for taking part in a Distarnet community is enough to loose all the digital data. This is the main reason why "eternal" media so attractive to the archival community. Even if an archive is not managed for a longer period of time, the assets are still ready to be used. However, we consider "eternal" media only to be an intermediate solution. "Eternal" media is either very expensive to create (e.g. rosetta disk) and not suitable for large amounts of data, or it still requires proper storage and handling. It might be a complement to digital archival, but it will not be a replacement. In many cases - especially where continuous funding can be expected an Distarnet only based solution might be preferable.

A shortcoming of all digital long-term archival concepts is that they require the active participation of the involved institutions. The digital data is passive and being acted on. However, mother nature itself shows as a concept of active data migration.

The Biological Model

In a simplistic model, living beings can be considered as containers carrying digital data defining the species. From this point of view, the individual being is of no value, only the survival of the species matters. The definition of the species is carried "digitally" using a two-bit code - the DNA. In this model, the living beings can be considered as the media into which the digital information is encoded. As we all know, living beings have a very limited lifespan. Therefore a regular migration is necessary. In nature, this migration process is achieved by the natural reproduction. In fact, Nature created a data storage medium which does the necessary migration automatically on its own[§]. This method has been highly successful. There are species - called living fossils - which exist on earth since more than 60 mio. years and the information defining these species has undergone many millions of migrations.

An even simpler, but as we all know, also a highly efficient and successful model is implemented by biological viruses: a virus (e.g. a flu virus) can be considered as a box containing a strip of information encoded as DNA. The DNA strip contains all the information about how to build the virus. In order to reproduce, the virus docks to a cell of living being and injects the encoded information into the cell. The cell integrates this information into it's own DNA, is reprogrammed and starts immediately producing viruses of the given type. In this sense, viruses are the perfect parasitic organisms. The only aim that a virus has seems to reproduce (migrate) and spread its genetic code.

These examples based on living beings led us to the idea that we should find a method to "teach" digital data to migrate itself on its own and to spread. Whereas using biological viruses or living beings as storage media for digital data is science fiction for the foreseeable future, there are "viruses" and "worms" well known to all of us which besiege our PC's constantly.

Artificial Life

To some extent, computer viruses and computer worms can be considered as "artificial life", as Eugene H. Spafford stated as early as 1994[**?**]. *Viruses* are small segments of computer programs which are attached to another program (host). If the host program is executed, also the code of the virus is executed. Usually, in a first step, the virus tries to infect other programs on the machine by attaching it's code to the original program code. In an optional, second step, the virus executes it's payload which might be a program that draws a funny icon on the screen (best case) or wipes out the disk (worse case). Viruses are spread by distributing infected programs to other computers (file sharing, floppy disk etc.).

Worms are defined as programs which can run independently and which spread and replicate using network connections. Worms often spread exponentially and can lead to serious degradation of network performance. Worms also can have a payload which is triggered by certain conditions and can harm the infected computers.

However, the concept of viruses and worms could be used for archival purposes. Instead of designing artificial life which acts like a pest - malware - it would be possible to carefully design autonomous programs (we will call them *archivlets*) which replicate and spread in a controlled manner, and which, as payload, carry our valuable digital data which has to be archived. Thus, the same principle that Nature uses to archive the information of species could be used for archival purposes.

Conclusion

While it is possible to achieve long-term archival of digital data with traditional means, we believe that the very special nature of digital data as described above combined with rapid obsolescence of information technologies requires a fundamentally different approach to archiving. Traditional migration concepts are very cumbersome, expensive and hazardous (as for example the money to do the migration must be available at the very moment the migration becomes necessary due to the technology change). The new concept is that the digital data should not be passively migrated, but it should *actively* migrate itself.

File sharing networks as the Distarnet present one of the alternatives for a less cumbersome migration method. Distarnet is based on currently available hardware and software technology. Yet, Nature itself shows very successfully that other methods of active data migration is possible. It has developed methods for archiving the information defining species based on a "digital" code – the DNA. In the domain of information technology, similar exists in the form of viruses and worms. Viruses and worms are small programs which replicate and distribute themselves without human interaction and often even against the will of the computer owners - on the Internet. Potentially these programs can carry arbitrary data with them (called "payload") and disseminate this data on the Internet. Because virtually all of these programs have undesirable or even damaging payload and negatively affect the infected computers, they are called "malware". However it is possible to use the same principle of replication and dissemination for archival purposes. It seems possible to find new concepts for long-term archiving based on biological models and Artificial Life. *Archivlets* which carry archival data as payload, self-replicating, self-spreading and adapting evolutionary to new technologies may be regarded as science fiction today, but are within the technological possibilities right now. With the ever growing capacity of the Internet both in speed and storage space, even large amount of data could be archived this way in a foreseeable future. The Internet can be regarded as a global organism consisting of billions of cells (=computers) hosting autonomous, active information particles which carry valuable information. Such a concept would eliminate the necessity of migration totally. Even if by todays standards such a concept seems to be far out, it is necessary to explore novel ways of digital long term archival. If we don't, the infamous "digital dark age" will become a sad reality.

References

[*] which poses large problems for the recording and movie industry. Audio-CD's and movie DVD's can be copied and distributed without loss and very little effort. The entertainment industry is desperately seeking methods (copy protection schemes) to prevent unauthorized copying.

[†] Photographic film material is highly non-linear and the interpretation of data read can be very difficult. Within the project state of the art photographic film material (and color microfilm) will be researched in terms of digital resolution, quantization depth, SNR, non-linear aspects like side-absorption and optical and chemical diffusion. To eliminate this hurdle the film is virtually linearized by applying appropriate compensation on the data in advance of the exposure.

[‡] thanks to Grant SNF #105714 of the Swiss National Science Foundation

§ Of course this model is over-simplistic since it does not take into account the evolution process. Evolution changes the data that defines a species gradually to adapt to changes in environment and can be considered as a gradual optimization

1. Bram Cohen, Bittorrent, http://bittorrent.com (2005).
2. IBM Zurich Research Labs, Ibm millipede, http://www.zurich.ibm.com/st/storage/millipede.html (2005).
3. Gordon E. Moore, Cramming more components onto integrated circuits, Electronics, 38(8) (1965).
4. The Long Now Foundation, Rosetta project, http://www.therosettaproject.org (2005).
5. Rudolf Gschwind Lukas Rosenthaler, Distarnet a distributed archival network, in Proc. IS&T's 2004 Archiving Conference, pp. 242 – 248 (2004).
6. Eurgene H. Spafford, Computer viruses as artificial life, Artificial Life, 1(3), 249 (1994).

Biography

Dr. Lukas Rosenthaler, born 1960, studied Physics, Mathematics and Astronomy at the University of Basel, and got 1987 a Ph.D. in Applied Physics in the field of Nanotechnology building an early Scanning Tunneling Microscope. During his Ph.D. thesis he got involved with image processing and image analysis. From 1988 to 1992 he worked as postdoc at the Swiss Institute of Technology in Zürich in an interdisciplinary project about the understanding and the computational simulation of the vision system of human beings. From 1992 to 2001 he worked in the field of computer graphics and visualization within Cadwork Corp., the leading software manufacturer for CADsoftware in Switzerland. During this period he also developed new methods for the restoration of damaged movie films, in affiliation with the Scientific Photography Lab of the University of Basel. Since 2001 Lukas Rosenthaler is a full time staff member of the Imaging and Media Lab of the University of Basel, Switzerland. The main research topics are the restoration of movie films and the long-term preservation of digital images. He also leads the project team of Distarnet.