Passive Digital Preservation Now & Later

Microfilm, Micr'Olonys and DNA

Vincent Joguin

Eupalia
France
vincent.joguin@eupalia.com
https://orcid.org/0000-0003-0627-8778

Abstract – This paper presents the newly available Micr'Olonys software-on-film digital archiving solution that builds on a multi-layered virtual machine, and microfilm, a proven medium with a life expectancy of 500 years. The passive digital preservation strategy it follows matches the features of a written form contrasting with current active strategies that rather compare to orality. Micr'Olonys stores digital files as 2D barcodes whose content can be restored, using any general-purpose computer and scanner, by implementing the procedure described in a self-contained bootstrap that spans a dozen human-readable pages including a simple 2-page algorithm.

In the future, Micr'Olonys aims to become a powerful preservation tool for complex formats and software, as well as the access key to DNA digital storage to archive massive amounts of data.

Keywords – passive, software, preservation, microfilm, DNA

Conference Topics – Designing and Delivering Sustainable Digital Preservation; The Cutting Edge: Technical Infrastructure and Implementation

I. Introduction

Digital contents and processing now permeate all human activities. Whereas natural languages take distinct oral and written forms, that of digital manifestations, mediated by computer hardware and software, is unclear. However, digital shares many characteristics with an oral form, with tremendously quick manipulation and transmission, fragile short-lived storage media, and constantly evolving environments and formats. As the Encyclopedia Britannica notes, "both [the Internet and oral tradition] work through rule-governed processes rather than fossilized texts; and both ultimately derive their strength from their ability to change and adapt."[1] Digital lacking to fully qualify as a written form, digital preservation is currently a matter of perpetually migrating media and contents, akin to mechanisms of oral tradition. The latter, whereby information is being preserved and transmitted orally, is characterized by the absence of any fixed form. Nevertheless, in oral cultures, accurate transmission of information may recourse to rhythmic speech (e.g. repetition, alliteration and assonance) which serves a purpose comparable to that of cryptographic hash functions or representation information used in digital preservation to maintain integrity and authenticity.

Indeed, writing digitally requires a series of conditions, from robust media to ensured long term readability. Until now, these conditions have never been met simultaneously, preventing a digital preservation approach similar to traditional paper archiving to emerge.

While oral tradition depends on people to ensure the continued transmission of information, migration-based digital preservation additionally depends on the continued availability of advanced digital technology and sustainability of energy production. However, in the current times of economic uncertainty and global warming, such resource-demanding preservation strategies are at risk of failing to pass our digital heritage on to future generations, which may have far-reaching adverse implications. Or, as the Latin proverb went, "*verba volant, scripta manent*": speech flies away, writing remains.

[1] https://www.britannica.com/topic/oral-tradition

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

DOI: 10.1145/nnnnnnn.nnnnnnn

II. Digital Preservation Strategies

From the inception of digital preservation in the early 1990s, a dilemma has grown about preservation strategies between migration and emulation. While migration – referring in this context to file format conversions over time – appeared as a simpler strategy technology-wise, emulation promised better authenticity and reduced costs. However, they were technological approaches rather than core digital preservation strategies.

Ultimately, only one core strategy came forth, that consisting of "the active management and maintenance of digital objects [...] so they can be accessed and used by future users."[2] Although this definition of Digital Preservation by the United Nations implies a migration approach as it focuses on digital objects, a similar definition implying an emulation approach could be formulated for active digital preservation focusing on digital environments. Active digital preservation strategies not only ensure that digital material can be accessed and used by future users, but more specifically that it is kept alive in future digital environments and can be smoothly accessed in the same or equivalent conditions as then current digital material. Such strategies are obviously desirable in many cases when the material is used frequently, when quick access to it is critical, and/or when it needs to be changed or combined with other material.

But what of digital material that is used infrequently over time, or most often never at all, and which may satisfactorily not be accessed immediately? This kind of material, usually archival records, arguably represents a very large portion of digital material that needs to be kept for the longest periods of time; active strategies, which are costly and prone to both economic and technological risks, do not seem appropriate for such material.

Therefore, a different strategy, passive digital preservation, could better fit the requirements to keep relevant digital material unchanged over time, in fact as is the case when preserving material on paper. Advancing the principles of the emulation approach, passive digital preservation builds on an immutable digital environment, whereas active digital preservation focuses on preserving significant properties in a constantly evolving environment.

While the challenge of active preservation is to accommodate to an evolving environment, that of passive preservation is to comprehensively capture an environment which is always very complex in the digital realm.

III. Reducing the Hardware Environment

A. *The Virtualization Approach*

From 1992, and for 20 years, Jeff Rothenberg has consistently formalized and supported a digital preservation approach focused on the digital environment[3]. His proposed approach [1] has led to the experimental development of a Universal Virtual Computer (UVC) by IBM, in collaboration with the Royal Library and National Archives of the Netherlands [2], which has since been further investigated [3].

Other projects have more recently emerged, both with the very similar approach of a simple dedicated virtual machine such as the iVM project[4] or [4], or using more conventional virtualization or emulation technologies such as the Olive project[5] supported by Vint Cerf who advocated the "digital vellum" concept [5], or the EaaSI project[6]. Some larger-scope initiatives also include similar approaches, in particular the PERSIST program of the UNESCO[7] and the BRASS program initiated by the DARPA[8].

B. *The Olonys Virtual Machine*

We started designing and developing the Olonys virtual machine at Eupalia from October 2000 until mid-2003. Work on the virtual

[2] https://archives.un.org/content/digital-preservation

[3] http://jeffrothenberg.org/Prof/digilong.html https://www.slideshare.net/FuturePerfect_/jeff-rothenberg-digital-preservation-perspective

[4] http://www.dlmforum.eu/jdownloads/eventsarchive/2019Bern/session2/2_1_Rummelhoff_%20stvold_Liab%20_Immortal%20virtual%20machine.pdf

[5] https://olivearchive.org/

[6] https://www.softwarepreservationnetwork.org/eaasi/

[7] https://unescopersist.com/

[8] S. Neema et al., Building Resource Adaptive Software Systems (BRASS), 2015-2019. https://www.darpa.mil/program/building-resource-adaptive-software-systems

machine was resumed in February 2009 within the context of the KEEP project[9].

The main distinctive characteristic of this virtual machine is to replace most of the hardware complexity with software, thereby drastically reducing the hardware environment necessary to instantiate the complete digital environment and to render associated digital content. Olonys achieves this by building on 5 different virtual processor layers, from a full-featured processor on par with the x86-64 or ARM 64-bit processors, down to a very simple processor that supports only two different instruction types (minimal processors are studied in [6]). Compared to IBM's UVC, the middle layer of Olonys (3rd layer) is of comparable complexity with 23 instructions, whereas the simplest layer is almost two orders of magnitude simpler, making it both much more robust, i.e., much less bug-prone, as well as easier and faster to implement.

IV. Materializing the Incorporeal

Reducing the reliance of software on hardware logic gates does not remove the need for software and digital content to reside on a very material substrate, and not in thin air, in clouds, as is still too often believed.

Common recordable digital media – magnetic such as hard disk drives and tapes, optical such as recordable CDs and Blu-ray Disks, electronic such as flash drives – are either not durable, or at least their durability has not sufficiently been proven, especially out of experience since they are all relatively new. Therefore, various digital media, usually optical[10], have been developed with a defined goal of durability. However, these media are either too expensive (e.g. when using glass) or have not been demonstrated to be more durable than common media (see for example [7]).

Existing analog media used digitally currently stand as the only alternative. Paper and film in particular are trusted carriers, especially when they conform to dedicated standards, for instance permanent paper defined by ISO 9706[11] and LE-500 rated microfilm[12], both of which being designed to last typically 500 years or more when stored appropriately. Moreover, film is both inexpensive and dense when compared with paper, while a magnifying glass is sufficient to retrieve filmed pictures.

Using such optical analog media digitally requires to convert bit streams into visual signals printed as pictures. Beyond the common QR code, Data Matrix and similar two-dimensional barcode standards, dedicated solutions have been developed for both paper (e.g. Optar[13], PaperBack[14] and PaperDisk[15]) and film (e.g. Monolith[16] presented in [8], PreserveOn[17] and Piql[18]). An approach tailored for one medium may also be used with the other, as exemplified by [9].

V. Optical Digital Bootstrap

While bits-on-paper and bits-on-film solutions overcome the barriers of media durability and independence from a specific technological device to access the preserved signal, a last hurdle remains: ensuring easy decoding of the signal back to exploitable information over the long term, when the necessary hardware and software stack will have long become obsolete and disappeared.

Based on the Olonys virtual machine presented previously, we have created a self-contained bootstrap that is part of Micr'Olonys, a software-on-film solution for long term digital preservation on microfilm. It is meant to be printed at least once per microfilm roll or

[9] Keeping Emulation Environments Portable (KEEP), FP7 project, 2009-2012.
https://cordis.europa.eu/project/rcn/89496/factsheet/en

[10] See for example:
http://www.arnano.fr/
http://www.glazt.com/
http://www.mdisc.com/
http://www.datatresordisc.eu/english/introduction-page-dtd.html
http://group47.com/
https://www.southampton.ac.uk/news/2016/02/5d-data-storage-update.page

[11] https://www.iso.org/standard/17562.html

[12] https://www.epminc.com/support/tech-tips/epm-product-information/21-tech-tip-050-life-expectancy-of-microfilm

[13] http://ronja.twibright.com/optar/

[14] http://ollydbg.de/Paperbak/index.html

[15] http://www.paperdisk.com/

[16] https://www.bitsave.ch/eng/index.html

[17] https://web.archive.org/web/20170602102100/http://preserveon.com/

[18] https://www.piql.com/

microfiche so that accessing the preserved content relies on no external resource other than a film scanner and general-purpose (i.e., Turing-complete[19]) computer, both arguably expected to remain readily available over the long term.

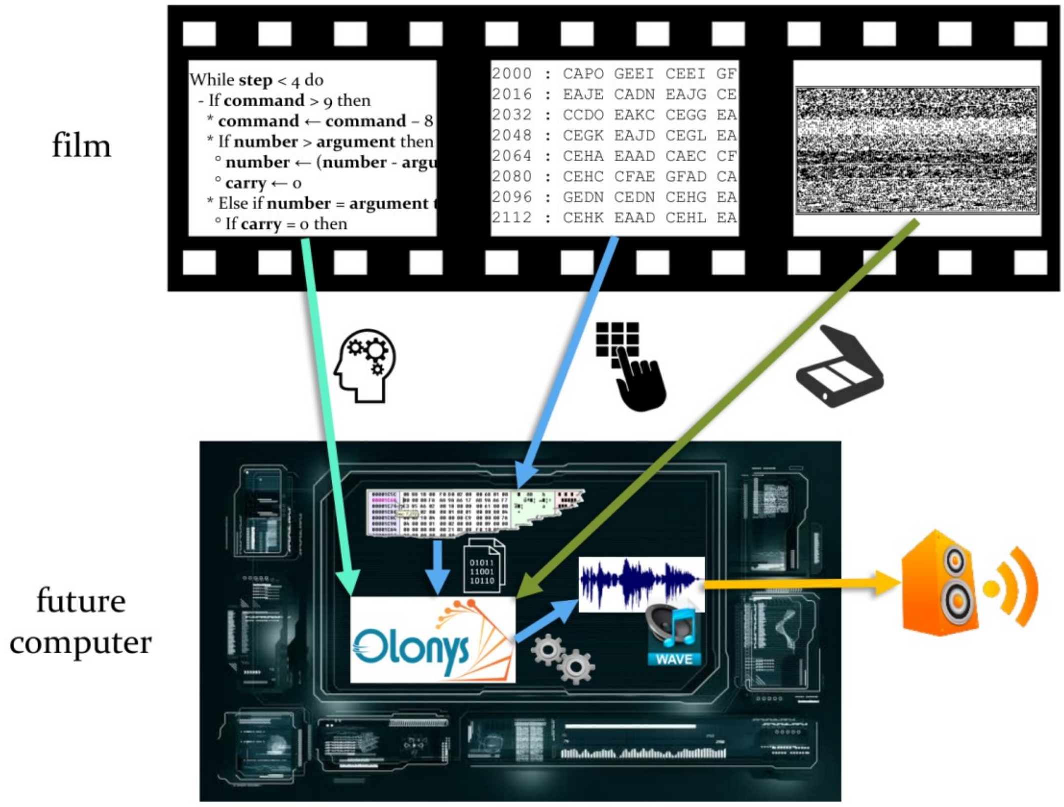
This optical digital bootstrap is capable of decoding dense custom two-dimensional barcodes with both intra-frame and inter-frame Reed-Solomon error correction. Using intra-frame correction, the original data will be entirely recovered although up to 7.5% of each barcode may be damaged. Inter-frame correction is able to completely rebuild 3 out of 20 heavily damaged or missing data blocks distributed within 20 different barcodes. This means that some barcodes may have suffered more than 7.5% damage, or up to 3 barcodes per series of 20 may be missing altogether.

The core of the bootstrap consists of 7 pages of worded technical instructions to precisely guide the user of the distant future through the process of setting up the environment necessary for program execution and proper scanning of the barcodes. It includes a simple 2-page algorithm whose implementation is responsible for execution of all necessary software, raw scanned images input and decoded file output.

A user-defined front page, 4 pages of listed letters to be input manually or using OCR upon initial execution of the algorithm, and 5 pages containing system barcodes complete the bootstrap for a total of 17 pages. The system barcodes contain internal Olonys software, including a dynamic recompiler that accelerates execution by a factor of more than 15, and the decoder for the more complex and advanced barcodes that hold preserved data.

Fig. 1 depicts the example of how the bootstrap is used to restore a sound file preserved on film.

The bootstrap is designed to rely on as few and as simple concepts as possible so as to minimize misunderstandings arising from changes over the course of decades or centuries to the natural language and concepts employed. The bootstrap is currently available in English and French, and both versions may be printed on the same carrier to help recover the intended meaning, similarly to how the Rosetta Stone was key in deciphering ancient Egyptian hieroglyphs.



```
While step < 4 do
  - If command > 9 then
    * command ← command – 8
    * If number > argument then
      ° number ← (number - argu
      ° carry ← 0
    * Else if number = argument
      ° If carry = 0 then
```

```
2000 : CAPO GEEI CEEI GF
2016 : EAJE CADN EAJG CE
2032 : CCDO EAKC CEGG EA
2048 : CEGK EAJD CEGL EA
2064 : CEHA EAAD CAEC CF
2080 : CEHC CFAE GFAD CA
2096 : GEDN CEDN CEHG EA
2112 : CEHK EAAD CEHL EA
```

Figure 1 Restoring a sound file preserved on film using the Micr'Olonys bootstrap

[19] Incidentally, Turing-completeness is the one property of computational systems that makes emulation and virtualization possible

VI. The Future: Software Preservation and DNA Storage

A. Preserving Complex, Dynamic and Interrelated Content

Although the Micr'Olonys solution embeds a general-purpose software processor used to automatically decode 2D barcodes, it is currently only able to restore files: their formats therefore need to remain sufficiently simple or normalized to avoid the problem of file format obsolescence. Recommended formats include BMP, uncompressed TIFF, WAV and ASCII text.

Future developments of the solution will include native support for more complex, compressed and composite formats such as PDF, JPEG2000 and 3D models. The bootstrap would be able to restore either the preserved file in its original format, or a conversion to an easily-exploitable uncompressed image or sound stream. Interactive and connected content such as dynamic websites, video games or software-based artworks are also planned to be supported using an additional self-adaptive peripheral management system.

B. Ultra-high Capacity Passive Digital Preservation

A major drawback of preserving digital data on microfilm is the relatively low capacity of this medium when compared with current digital storage technologies. For instance, a movie soundtrack can typically fit on digital microfilm, whereas preserving movie frames in lossless quality would require many more reels than an analog film shoot, which is simply impractical.

To qualify for passive digital preservation, any carrier with higher capacity should however fulfill the properties of passive storage and independence from a specific technological device for access. DNA storage matches these criteria: as a natural information carrier, it is proven to preserve information over millennia when stored in appropriate conditions; the ability to sequence DNA, although a highly complex process, can be expected to endure for as long as interest in DNA will exist, i.e., arguably over the very long term. Decoding digital information preserved within DNA will however never be trivial, and therefore an external, simpler carrier will need to be used as an access key. Microfilm or paper would typically be used in this context.

Many actors, including the Paris-based DNA Script company[20] and Microsoft[21], are collaborating to make fast and affordable DNA writing a reality. Significant investments are needed to complete the research and development necessary to reach an operational solution, but the incentive now exists to support the effort. The R&D process is expected to last for another decade and to progress similarly to the process that led to DNA sequencing, i.e., at a pace faster than Moore's law. Simultaneously, digital preservation approaches dedicated to DNA have started to emerge (see for example [10]). Conforming to acceptable ethics standards may ultimately constitute the main challenge in DNA storage adoption.

If DNA storage becomes a reality, it would stand as an extremely competitive solution to preserve digital information for which immediate access is not required and/or as a fail-safe complement to active preservation solutions.

Acknowledgment

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 231954, as well as from Bpifrance, and is currently being supported by the technical industries financial instrument of the Centre national du cinéma et de l'image animée (CNC).

References

[1] J. Rothenberg, "Ensuring the Longevity of Digital Information," 1999. http://www.clir.org/pubs/archives/ensuring.pdf

[2] R. Lorie and R. van Diessen, "UVC: A Universal Virtual Computer for Long-term Preservation of Digital Information," 2005. https://domino.research.ibm.com/library/cyberdig.nsf/0/10229b6de0d054c585256fa900681995?OpenDocument&

[3] N. Krebs, L. Schmitz and U. Borghoff, "Implementing the Universal Virtual Computer", *EUROCAST 2011*, in Lecture Notes in Computer Science, vol. 6927, pp. 153-160, Berlin, Heidelberg: Springer, 2012.

[20] http://dnascript.co/

[21] https://www.microsoft.com/en-us/research/project/dna-storage/

https://www.researchgate.net/publication/310576462_Implementing_the_Universal_Virtual_Computer

[4] L. T. Nguyen and A. Kay, "The Cuneiform Tablets of 2015," *SPLASH 2015*, Oct 2015, Pittsburgh, PA. http://www.vpri.org/pdf/tr2015004_cuneiform.pdf

[5] V. Cerf, "Digital Vellum and Archives," *FASTER CoP*, Aug 2016. https://www.nitrd.gov/nitrdgroups/index.php?title=DigitalVellumAndArchives

[6] W. F. Gilreath and P. A. Laplante, *Computer Architecture: A Minimalist Perspective*, The Springer International Series in Engineering and Computer Science, vol. 730, Boston, MA: Springer, 2003.

[7] J. Perdereau, "Recherche sur la conservation des archives numériques dans le cadre du GIS DON," 2012, page 9. http://www.lne.fr/publications/guides-documents-techniques/duree-vie-DVDR-DVD-SYYLEX.pdf

[8] P. Fornaro, R. Gschwind, L. Rosenthaler and A. Wassmer, "Monolith: materialised bits, the digital Rosetta film," *DH 2014*, Jul 2014, Lausanne, Switzerland. http://dharchive.org/paper/DH2014/Paper-615.xml

[9] S. W. Schilke and A. Rauber, "Long-term archiving of digital data on microfilm," *Int. J. Electronic Governance*, vol. 3, no. 3, pp. 237–253, 2010. https://publik.tuwien.ac.at/files/PubDat_191964.pdf

[10] M. Dimopoulou, M. Antonini, P. Barbry and R. Appuswamy, "A biologically constrained encoding solution for long-term storage of images onto synthetic DNA," *EUSIPCO 2019*, Sep 2019, A Coruña, Spain. https://hal.archives-ouvertes.fr/hal-02064896