EPJ Web of Conferences **214**, 09003 (2019) https://doi.org/10.1051/epjconf/201921409003
CHEP 2018

The obsolescence of Information and Information Systems

CERN Digital Memory project

Jean-Yves Le Meur[1*] and *Nicola* Tarocco[1]

[1]CERN

Abstract. In 2016 was started the CERN Digital Memory project with the main goal of preventing loss of historical content produced by the organisation. The first step of the project was targeted to address the risk of deterioration of the most vulnerable materials, mostly the multimedia assets created in analogue formats from 1954 to the late 1990's, like still and moving images kept on magnetic carriers. In parallel was studied today's best practices to guarantee a long life to digital content, either born digital or resulting from a digitization process. If traditional archives and libraries have grown up during centuries establishing recognized standards to deal with the preservation of printed content, the field of digital archiving is in its infancy.
This paper shortly exposes the challenges when migrating hundreds of thousands of audio, slides, negatives, videotapes or films from the analogue to the digital era. It will then describe how a Digital Memory platform is being built, conform to the principles of the ISO-16363 digital object management norm that defines trustworthy digital repositories. Finally, as all information repository managers are faced with the necessary migration of underlying systems and the obsolescence of the information itself, the talk will explain how a digital archiving platform focusing only on content preservation could be of direct interest for most of the live systems.

1 Introduction

About four thousand years ago in Crete, a disc of clay was covered with hieroglyphs using 241 spiral-shaped signs printed with punches. This disc, known as the Disc of Phaistos, remains a riddle today: we know with certainty neither its meaning, nor its use, nor its place of manufacture. No other similar discs were found despite the use of punches that suggest the existence of an alphabet. Its authenticity is sometimes disputed and its dating is imprecise.

Thus, the obsolescence of information and information systems is not new and several similar examples illustrate the difficulties of transmission in the pre-digital world. But with the advent of the digital age after the invention of the World Wide Web, it took less than 30 years to observe the speed of information obsolescence in parallel with the explosion in the amount of information. The famous "404 Error" is certainly the most popular page on the web, as evidenced by numerous studies on both current web pages [1] and academic production [2]. Digital obsolescence includes many origins other than link-rot, for example:

- Bit rot: the alteration of the electrical charge of one bit on a RAM.

*e-mail: Jean-Yves.Le.Meur@cern.ch

- Technical Obsolescence: the player no longer allows the media to be played back.
- Format obsolescence: the format of the file has become unknown; figures or fonts are no longer interpreted correctly.
- Migrations and transitions: systems managers change regularly (every 2 to 20 years), software every 5 to 10 years and hardware every 3 to 5 years; each change entails a risk of data loss.
- Dissipation: data poorly described has become unavailable, lost in cyberspace.
- Ambiguous intellectual property: the lack of source information makes it impossible to know if a fake or altered copy of the original data is being processed.
- Computer attacks and Human errors, not always immediately detected.

Awareness of this phenomenon is increasingly shared. Beyond the experts in virtual archiving, the main actors of ICT have already acknowledged the existence of a major risk that the human memory will be unable to trace the beginnings of the "digital age". Vint Cerf, vice President of Google, warned against an "information black hole" and the need to create a "digital vellum" [3].

2 Digital preservation initiatives at CERN

- **2004: LTEA's recommendations**

  As early as 1997, the CERN set up a Working Group on Long-Term Electronic Archiving (LTEA) [4] with a mandate to identify the types of electronic documents created by the institute and to develop an archiving policy. Recommendations were issued and approved in 2004, including these:

  (a) implement selective email archiving

  (b) continue archiving the official CERN Web with external partners

  (c) define and implement a management plan for all CERN documents

  (d) postpone operating a real digital archive until later, when it will appear less expensive

  (e) in the meantime, ensure no information is lost due to format migration or otherwise.

  These decisions contributed to the evolution observed in the 2000s, in particular through the consolidation and centralization of the organization's website and email management, as well as the archiving of group conversations and the outsourcing of public web archiving to the Internet Memory Foundation.
  Gradually, more and more types of documents to be preserved converged into CERN's two main information systems, the Document Server (CDS) and the Engineering Equipment Data Management Service (EDMS). The digital preservation of the official institutional documents was delegated de facto to these systems, considered as "Certified Information Systems".

- **2009: the High Energy Physics Data Preservation Project**

  Experiments in High Energy Physics (HEP) can hardly be reproduced and the data generated are therefore unique. Under the impulse of DESY, the DPHEP project has undertaken to coordinate the efforts of most of HEP institutes to ensure data preservation according to the FAIR principle: findable, accessible, interoperable and reproducible [6].

In 2012/2013, the European Strategy for Particle Physics (ESPP) proposed strategies for the preservation of physics data. These helped as guidelines for the project. In addition to bits preservation that is being addressed as a mandate of the CERN Data Center, new activities have been launched to ensure additional preservation of the analyzes, documentation, software and associated environments, like the CERN Open Data service, the CERN Analysis Preservation portal and the Reproducible research data analysis platform (REANA) [5].

By fostering initiatives of the Experiments to establish ambitious preservation plans and throughout the developments of the related projects, the DPHEP project at CERN is moving in the direction of the ISO 16363 standard that defines "Trustworthy Digital Repositories" (TDR) [7]. The objective is clearly to move from "preservation by chance" practices to standardized and systematic good practices.

- **2016: CERN's Digital Memory**

  With the launch of the Digital Memory Project in 2016, CERN's IT department faced the challenge of preserving documents produced by the Organization in digital form with standard practices. An additional step must be taken to protect digital content against contingencies that inevitably lead to information loss. The implementation of an OAIS-compliant infrastructure (detailed in section 4) is envisaged on the basis of the Invenio software used by many document services.

  In parallel to the computing developments, "the big data of the 20th century" - the multimedia - must be treated in urgency: the content stored on analog carrier is threatened with obsolescence. The magnetic tapes and other media on which the CERN's audiovisual documentation has been recorded since 1954 are beginning to deteriorate. The conversion of this heritage into a digital format intends to preserve its content and open its access to a large audience.

3 Multimedia data in danger

3.1 The photographs

The intense activity of CERN's photography department, with nearly 400 requests for shootings per year, seems to vary both according to the major stages of the Organization (above the curve on the fig 1) and the developments in the photographic techniques used by the laboratory (below the curve). 120,000 black and white images taken between 1955

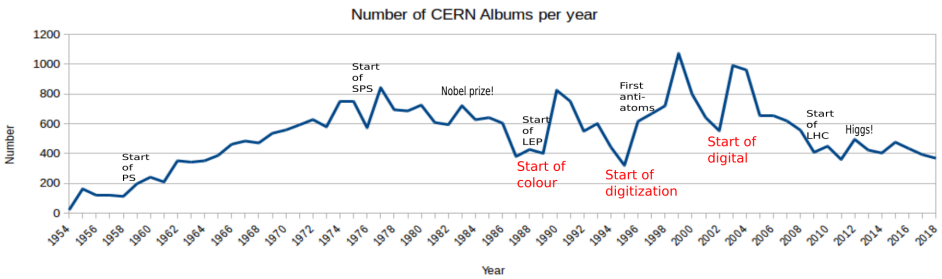

Figure 1. Evolution of the photographic collection based on an estimate of albums

and 1986 were scanned, grouped into albums and put online in 2014. About 300,000 colour images are being processed now. Medium and large format negatives, slides and positives are converted into TIFF files at 4800 ppi for a size of 24x36cm, and 8 or 16 bits/channel RGB depth depending on the collection. Multiple derived JPG versions of different widths are also produced to allow optimal dissemination. The main challenge now is to successfully enrich past albums with quality captions on individual images. The recognition of the individuals, places, technical objects or infrastructures photographed was made possible thanks to the participation of pensioners within CERN's Alumnis network. This challenging crowd-sourcing work has allowed to caption about a hundred pictures so far.

During this project, the discovery of hundreds of slides taken during the construction of the LEP but left in sub-optimal archiving conditions gave rise to a CERN art collection [8]. For years, microorganisms have proliferated by feeding on the emulsion that covers the

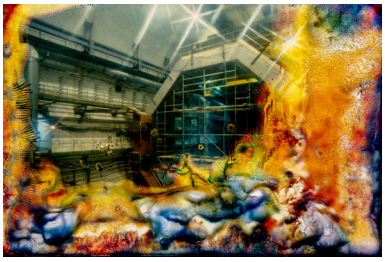

Figure 2. L3 detector slide partly eaten by mould

slides. To everyone's surprise, the projection of these organically modified images revealed works of undeniable aestheticism. A collection of CERN's digital memory art collection has resulted, as a reminder of the importance of preserving. It is now the subject of an exhibition in an art gallery in the old city of Geneva.

3.2 Videos and films

The risk of loss of audiovisual data on magnetic tape is important and their digitization was undertaken on a large scale in the 2000s by many institutions specialized in this field, such as the INA, TSR or IOC organizations.
At CERN, the video production unit has a substantial content with a total duration estimated at 6'200 hours and dating back to the 1960s (including the rushes "raw data"). Between 2002 and 2006, a fraction of the media was digitized for the purpose to make them accessible on the web, but without the intention of transferring their preservation from the analogue medium to a digital preservation format. Generally stored in bad conditions for long-term archiving,

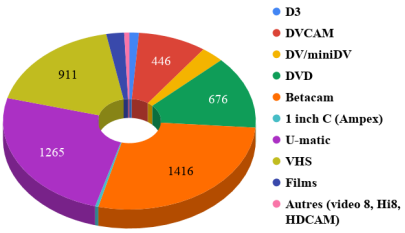

Figure 3. Distribution of audiovisual media after inventory

the media are very varied (see fig 3) and the metadata very limited. The inventory inherited

Table 1. Video file formats specifications

Originals	Preservation master	Production master	Access copy
Films (16/35/8 mm)	.mkv FFV1 10 bits RGB res. 2K or 4K	.mov Apple ProRes 422 HQ res. 2K or 4K	.mp4 H264 @ 5Mbps res. 1920x1080
Analogue and digital SD video	.mkv FFV1 10 bits YCbCr res. x576 height	.mov Apple ProRes 422 LT-SD res. x576 height	.mp4 H264 @ 1Mbps res. 640x360

from several generations of operators is not always complete nor consistent and the experts who could enrich the records are no longer available.
However, the urgency of the treatment imposes to make an essential choice as to the archiving format of the videos. Digital preservation standards have difficulty to emerge and are extensively discussed in the international audiovisual community (IASA) [9]. Following the recommendations of the Memoriav Foundation [10] which supported this project, three formats are created for each medium: a preservation file (.mkv), a production file (.mov) and an access file (.mp4) as shown in table 1.

3.3 Soundtracks

CERN committees have been recorded since 1956 for the purposes of the translation service. These are kept in good conditions in the store of the central archive of the organization. 8'400 recordings cover important historical meetings in the form of BASF cassettes and AGFA 1/4 inch tapes. All remain confidential. They are complemented by a small public collection of interviews with key figures in the organization.
They are being converted to a preservation file (uncompressed WAF, 16 bit 48 kHz), with derivative access files (.wav, .ogg, .mp3) after trimming, normalization, noise reduction and sometimes concatenation of several parts.

Once delivered by the digitization companies and quality-controlled at CERN, all the multimedia files will enter a new cycle. Exposed to the new risks of the digital obsolescence, these should now be properly preserved for the long term.

4 Towards a standard digital preservation solution

The relationship between a trusted digital archive (Open Archival Information System archive) and live digital information systems is equivalent to the classic relationship between an institutional paper archive and physical libraries. Within the digital paradigm, the three points below deserve special mention:

- All contextual information (hardware, software, related data structures, formats) useful for preserving content is available at the time of content creation but may be inaccessible a few years later.
- The later the implementation of an OAIS archive, the more expensive the transfer of existing data will be, if not impossible.
- Any new information system should include a digital archiving strategy or module upstream of its implementation. This requirement is gaining momentum around the world by becoming one of the mandatory criteria for funding new ICT projects (e.g. the German

Research Foundation, UK Research and Innovation or the Netherlands Organisation for Scientific Research).

4.1 The OAIS model

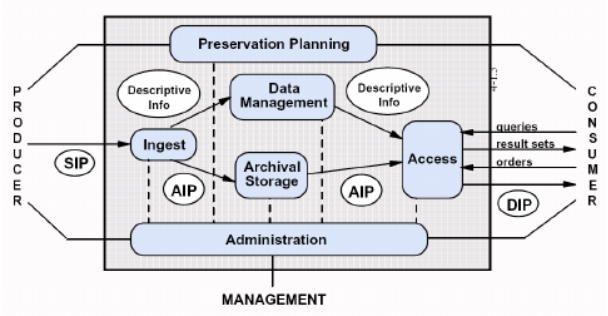

Figure 4. Open Archival Information System reference model

The OAIS reference model provides organizations with common guidelines for establishing preservation plans and an infrastructure that ensures "FAIR" preservation [11]. Many tools are available that are inspired by this model, often implemented on top of institutional repository software. The principle is based on three types of data:

1. SIP: the 'submission information package' is the entity created by the repository of the object (document in the broad sense), which must contain at least the intrinsic content of the object and the descriptive metadata.

2. AIP: The archival information package is the self-sufficient entity that guarantees the survival of the object in the long run, regardless of the information system in which the object is processed. This 'package' must contain the object itself (in a preservation format), the descriptive metadata of the object, all the information related to its preservation, the contextual information (provenance, rights, etc.) and information explaining the 'packaging'. The idea is that an AIP discovered in 4'000 years does not become a new "Disk of Phaistos".

3. DIP: the 'dissemination information package' is a transformation of the AIP intended for the dissemination of the object. It will often contain formats derived from the preservation format, and reduced metadata directly usable for easy consultation.

From this model is born the ISO 16363 standard which defines the certification rules of a trustworthy digital deposit (TDR) [7].

4.2 OAIS at CERN: E-Ternity

Evolving CERN Information Systems into Trustworthy Digital Repositories would still involve efforts at three levels for CERN:

1. Organizational: the commitment to support a digital archive must be clearly established at the managerial level, in the form of a policy or any other institutional rule, together with a preservation plan.

2. Infrastructure: the documentation of all data storage and backup processes (memory or science) must be complete and up-to-date.

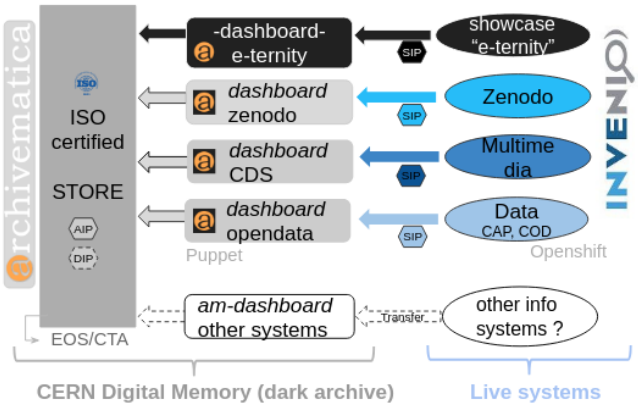

Figure 5. OAIS compliant Information Systems Architecture at CERN

3. Digital Object: Data that CERN commits to preserve numerically must be stored in the form of AIPs.

CERN's information systems are numerous. Invenio open source software supports a number of key systems such as the Institutional Repo (CDS), the HEP repo (Inspire), the Long tail of Science repo (Zenodo), the Open Data and Preservation Analyses services. By plugging Invenio with an OAIS archiving module, Invenio-based systems will be able to ensure standard digital data retention without too much investment.
Information systems are responsible for producing SIP packages for content that must be archived. These SIPs are transferred on a dedicated pipeline from each system to the archiving system. By using the Archivematica software, a specific dashboard can be set up for each pipeline, which makes it easy to configure the appropriate processes for different content as well as to partition the accesses in the case of restricted content. Information systems based on other software, such as Indico, Mailing lists, EDMS or others, can also be added to the permanent archive either by also creating SIPs absorbed via a specific dashboard, or by using a filesystem based transfer that delegates the creation of SIP to Archivematica micro-services.

A first prototype, named E-Ternity, has shown the feasibility and the difficulties of a solution of this type. The developed modules have been containerized so that as many dashboards as needed can be easily deployed in the CERN Data Center. Various dashboards have been configured to perform ingestion of data types requesting different modes of identification, conversions, fixity checks, etc. The main difficulty encountered is related to the size of the files to be kept, especially for audiovisual data and physics datasets. The handling of these files is both CPU and time consuming. In order to avoid redundant copies of one file system cluster to another one, the option to consider the object as an outside resource and to create AIPs composed only of metadata is considered. Of course, this requires that the external resource is maintained on a trusted system and that a monitoring of this system is in place at the level of the long term archive.
In this architecture, the OAIS archive is considered as a "dark archive". Dissemination is managed upstream by the live information systems. The possibility of generating DIPs remains open, if necessary. Only information system managers will actually have the ability to retrieve a piece of archived content in a readable format, as well as the complete preserved collection in case of a system migration, or a major disaster.

5 Conclusion

In these early years of the digital age, the transmission of knowledge to future generations requires extended awareness and strong commitments. With the on-going multimedia digitization taking place at CERN, some 550 TB of new files are being copied to the CERN Data Center, side by side with the 300 PB of the LHC physics data. The scale is different but the exposure to the new risks of the digital preservation are similar.
By coupling the Invenio and Archivematica software, the E-Ternity proof of concept is evaluating how CERN Digital Memory and HEP Data could benefit from a common layout to start storing content into an OAIS compliant archive.

References

[1] Van der Graaf, Hans. "The half-life of a link is two year", ZOMDir's blog. http://blog.zomdir.com/2017/10/the-half-life-of-link-is-two-year.html (2018)

[2] Habibzadeh, P. "Decay of References to Web sites in Articles Published in General Medical Journals: Mainstream vs Small Journals"; Sciences, Schattauer GmbH - Publishers for Medicine and Natural (2013-01-01). ; https://doi.org/10.4338/ACI-2013-07-RA-0055

[3] Cerf, Vint. The Guardian, http://www.theguardian.com/technology/2015/feb/13/google-boss-warns-forgotten-century-email-photos-vint-cerf.

[4] Pollermann, Bernd et al., Report on Long-Term Electronic Archiving (LTEA), 2000, http://cds.cern.ch/record/1028139

[5] Chen, Xiaoli et al. "Open is not enough", Nature Physics 2018, 1745-2481, https://doi.org/10.1038/s41567-018-0342-2

[6] DPHEP Collaboration. "Status Report of the DPHEP Collaboration: A Global Effort for Sustainable Data Preservation in High Energy Physics", arXiv:1512.02019v2 [hep-ex], 2017/02/17, http://doi.org/10.5281/zenodo.46158

[7] "Space data and information transfer systems – Audit and certification of trustworthy digital repositories," ISO Norm 16363:2012, http://www.iso.org/iso/catalogue_detail.htm?csnumber=56510

[8] Schaeffer, Anais. "Ready to get mould on your walls?", CERN Bulletin, 2018/04/13, http://cds.cern.ch/record/2315836

[9] Blood, George et al. "Guidelines for the Preservationof Video Recordings IASA-TC 06", 2018, https://www.iasa-web.org/sites/default/files/publications/IASA-TC_06-A_20180518.pdf

[10] Jarczyk, A et al. "L'archivage numérique des films et vidéos : fondements et orientations," tech. rep. Memoriav (2015), http://memoriav.ch/recommendations-dafv-en/

[11] Wilkinson, Mark D et al. "The FAIR Guiding Principles for scientific data management and stewardship", Nature Scientific Data, 2016/03/15, https://doi.org/10.1038/sdata.2016.18