The ARCHIVATOR – A Solution for Long-Term Archiving of Digital Information

Oscar Plata; University of Malaga and Tedial, SL; Malaga, Spain
Rune Bjerkestrand; Cinevation AS; Drammen, Norway

Abstract

Digital archiving and preservation is the set of processes to conserve digital data and ensure its accessibility for the long term. As digital objects are stored in some binary format, the above processes involve the ability to interpret the binary format and transform it to a form understandable by humans.

ARCHIVATOR industrial consortium was set up in mid-2009 with the aim of developing a reliable, secure, cost-effective long-term digital archive solution using photosensitive polyester, a very stable material that remains unchanged for a period of centuries in optimal conservation conditions.

Introduction

Digital archiving and preservation is the set of processes to conserve digital data and ensure its accessibility for the long term, which may be understood as 50 to 100 years or even longer [1]. As digital objects are stored in some binary format, the above processes involve the ability to interpret the binary format (physically stored) and transform it to a form understandable by humans (logical information).

When considering the preservation of digital data for the long term a variety of technical challenges come up [2][3]. These difficulties can be separated into two main areas, physical preservation and logical preservation. The first area refers to the reliability, affordability, sustainability and efficiency of the physical media used to archive data (storage media). The physical and chemical stability of the recordable medium where data is stored is in the core of the problem of long-term digital archiving [4]. The second area covers various levels. Near the physical level, we must ensure the ability to recover the bit stream from the physical media regardless of its possible degradation or the hardware obsolescence of the storage device. Above the physical level, we must also preserve the understanding and usability of the digital data despite changes in data formats and software applications used to read and interpret such formats (software obsolescence). Finally, information describing archived data (metadata) is required in order to ensure its integrity and authenticity.

These difficulties fostered the development of the ISO data reference model OAIS (Open Archival Information System), defined by recommendation of the CCSDS (Consultative Committee for Space Data Systems) and published in 2002 [5]. This model has become the standard for long-term digital preservation and focuses mainly on the second area discussed above. Essentially, it defines a framework to establish basic concepts and tasks required for the storage and accessibility of digital data for the long term.

The OAIS model was designed to protect digital data through systematic periodic data migration that involves not only the bits but the full information content (preservation description information). Data migration is the most widely preservation strategy today to address the technical obsolescence problem. This strategy is based on the property of digital data to be copied an unlimited number of times with very few errors, which is due, to a large extent, to the extensive use of codes for error detection and correction [6]. A drawback of this strategy is the risk of altering the digital data throughout the migration process, compromising its accuracy, integrity or completeness (data corruption). There is also the factor of costs associated to the periodic migration process.

An alternative strategy to data migration is emulation, an approach that preserves digital data in the original physical media but provides tools to survive hardware/software obsolescence. With emulation tools, old systems (applications, formats, device drivers …) can work on the new systems that will be available in the future. Archiving a description of such tools associated to the digital data in a stable storage media would allow to reduce or even eliminate data migration for the long term. While emulation has not been widely adopted to date for digital preservation, however there is a notable research interest on it [3][7].

Apart from migration or emulation, another important option to mitigate the impact of technology obsolescence rests on the use of standards for digital data/metadata formats together with open-source software for rendering such formats. The open technology is the best to be preserved, as full description of the format and source code of the software are accessible and can be archived with the data. However, this is not current practice to date in many industries. A notable example of an activity in this line is the LTR-TWG (Long Term Retention – Technical Working Group), formed in 2008 by the SNIA (Storage Networking Industry Association) [8] to address storage aspects of digital preservation. This working group proposed SIRF (Self-contained Information Retention Format) [9] in 2010 to define a standard storage format providing long term preservation functionality. SIRF offers encapsulation of data and associated metadata at the storage level.

In this context, the ARCHIVATOR industrial consortium was set up in mid-2009 with the aim of developing a reliable, secure, cost-effective long-term digital archive solution using non-fading, photosensitive polyester-based micrographic film, a very stable material that remains unchanged for a period of centuries in optimal conservation conditions.

The rest of the paper describes the activities carried out in ARCHIVATOR project to date, focusing on the technologies we are developing to put in the market a cost-effective industrial solution to the digital long-term archiving and preservation problem.

©2012 Society for Imaging Science and Technology

The ARCHIVATOR Project

The ARCHIVATOR project aims to build an infrastructure to conserve digitized content on high resolution micrographic film (long-term archiving). The process shall be integrated into short and medium-term digital storage workflows. The project was started in mid 2009 and is scheduled to finish by the end of 2012.

The ARCHIVATOR Consortium

The ARCHIVATOR consortium consists of four technology based European companies: Cinevation AS (Norway), which develops digital recording systems on micrographic film, In-Vision Digital Imaging GmbH (Austria), which designs professional optical systems, P+S Technik GmbH (Germany), which develops professional film equipment, including micrographic film scanners, and Tedial, SL (Spain), which designs software solutions for audiovisual and multimedia information management and archiving. In addition, the consortium includes two film post-production companies, representing the market of potential users: Nordisk Film Post Production AS (Norway) and CPAA (Spain). The activities of this consortium are being funded by the European program EUREKA Eurostars [10].

The ARCHIVATOR Consortium

Company	Who are they How do they contribute
CINEVATION AS (Norway) www.cinevation.net	Development and manufacturing of film recording and printing products and systems
	Lead project partner as well as project manager. Cinevation develops and provides technology and equipment for data recording.
IN-VISION Digital Imaging GmbH (Austria) www.in-vision.at	Company specializing in optical system design and manufacturing of sophisticated optical systems
	Develop and provide illumination and imaging optical systems for data recording and data scanning
P+S Technik GmbH (Germany) www.pstechnik.de	Cine equipment manufacturer
	P+S Technik develops and provides technology and equipment for data scanning
TEDIAL, S.L. (Spain) www.tedial.com	Software company that develops a complete solution for management of media contents
	Development of packaging and un-packaging of data, as well as integration with an IT infrastructure
Nordisk Film Post Production AS (Norway) www.nordiskfilm.com	Post production and film laboratory facility
	Represent user market: product development, marketing and application testing
Centro de Produccion Audiovisual Autor, S.R.L. (Spain) www.cataestudios.com	Post production facility and copyright authority in Spain
	Represent user market: product development, marketing and application testing

The ARCHIVATOR Solution

The ARCHIVATOR project is developing a complete solution to the problem of long-term storage of digital data ensuring its integrity. The solution consists of secure, reliable, and cost-effective long-term data archival system, leveraging the well documented archival properties of micrographic film to store digital data within an information-technology (IT) framework, where it is integrated like a any other storage system.

This project utilizes specialized non-fading polyester-based photo-sensitive high resolution micrographic film to archive all forms of digital data, including, but not limited to, documents, databases, high definition images, and fully mixed-down audio visual sequences.

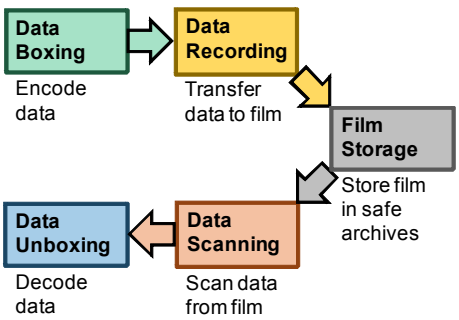
The potential of this technology extends across a broad sector of the market, including governmental institutions, conservatories, libraries, registers, insurance companies and banks, film studios, television broadcasters, scientific research bodies, aeronautical engineering agencies, oil and mining exploration permit holders, and medical researchers.

Workflow

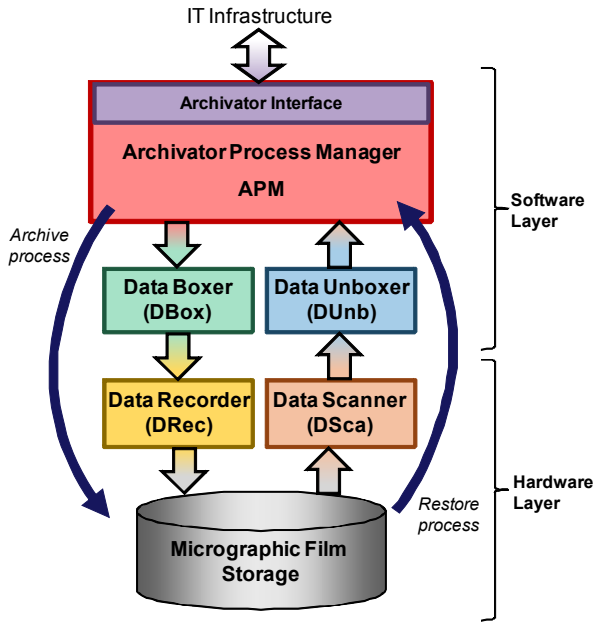
The figure below shows the workflow we used in our system. It is composed of four processes, together with the task of storing the recorded film. Each process is in charge of the following tasks:

- **Data Boxing**: The process of converting digital data so that it can be recorded into the film. This process includes ingestion, formatting, and encoding of data. Metadata is also added at this point so that the film record can be easily identified and tracked. Metadata is also added to ensure data preservation.
- **Data Recording**: The process of exposing the previously boxed data onto high resolution micrographic film. High-speed recording technology is used to get the job done quickly. We also focus on optimizing the data density and bit depth on the film to maximize the reproducibility of the recording created.
- **Data Scanning**: The process of regenerating the recorded data at high resolution. This step turns the data from a film-based dataset back to a computer readable digital dataset, thereby consolidating the process chain.
- **Data Unboxing**: The process that fully restores the scanned data so that it appears as it did in its original 'unboxed' form. This step involves decoding and re-formatting the data to make it fully readable.

The ARCHIVATOR workflow



The ARCHIVATOR structure



Structure

The ARCHIVATOR solution consists of a stable and durable storage media and a software layer able to encode data properly and to integrate the system into a storage infrastructure.

The figure above shows the internal structure of the archival system, comprising the four modules discussed in the previous section. These modules are in charge of implementing the two basic processes: archive process and restore process. The archive process takes data (digital files) from the outside world and, after applying some formatting operations, stores it in the micrographic film media. The restore process decodes the recorded data and put the recovered data files in some place in the outside world.

The structure of the system can be split into two layers, one hardware and one software. The hardware layer comprises the recorder and the scanner of the micrographic film storage media. This two modules are designed for high-resolution, high-speed recording/scanning in order to reach high-density recorded data.

The organization of the software layer is inspired by the OAIS model, with two basic modules in charge of ingesting (Data Boxer) and recovering (Data Unboxer) data. These two modules are in turn integrated into a process manager (APM) that automates the operation flows required to store and retrieve data in the system.

The Data Boxer is, in turn, composed of two sub-modules, one devoted to error correction and the other to the formatting operation. The first sub-module is in charge of including forward error correction (FEC) codes to the input data file. These codes represent some data redundancy that is used by the data unboxer during the recovering process, in order to detect and correct recording and/or scanning errors. The second sub-module takes care of formatting the bit stream data, with the FEC codes already applied, into a sequence of image frames suitable to be recorded into the film media.

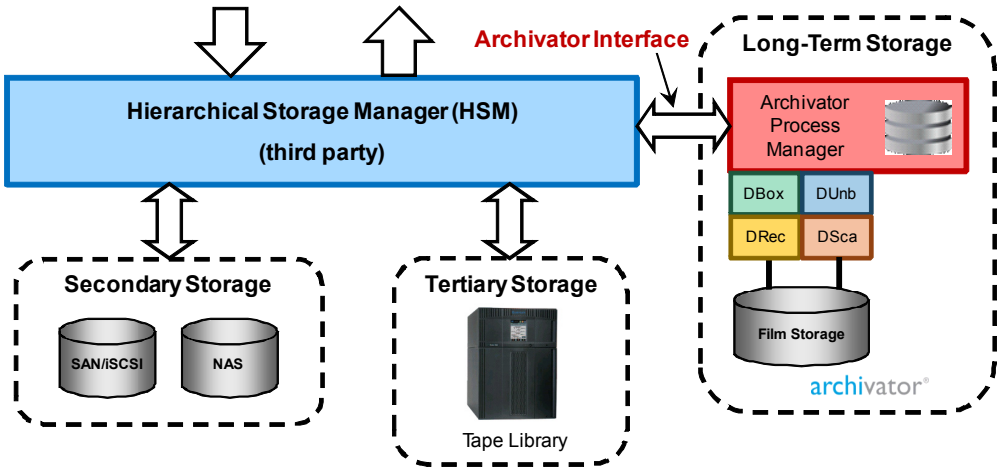
The Data Unboxer, on the other hand, is also composed of two similar sub-modules that carry out the inverse operations. The first sub-module interprets the scanned image frames and transforms them into binary data, with the FEC codes applied by the Data Boxer still included. After this process, the second sub-module takes into account such FEC codes and uses them to accomplish the error correction task and recover the data object.

The Archivator Process Manager (APM) is a software system in charge of defining, planning and monitoring data transfers between the outside world and the data recorder and scanner through the boxing and unboxing processes. APM offers a web services based interface to the outside world that can be used to initiate boxing-recording (archive), scanning-unboxing (restore) and recording verification processes.

IT Infrastructure Integration

The Archivator Process Manager also permits a seamless integration of the micrographic film based archival system within an overall IT infrastructure through a hierarchical storage manager (HSM), as shown in the figure below. Our vision is a system

Integration of ARCHIVATOR into an IT infrastructure



©2012 Society for Imaging Science and Technology

similar to those currently used in, for instance, the television broadcast industry. Digital media archives are classified as either 'on-line', 'near-line', or 'off-line', and the corresponding data is then managed accordingly, using different storage media for the different levels of storage classification.

The goal of the ARCHIVATOR system is that, with a simple click of their mouse, users will be able to classify data for the different types of archival systems, and automatically initiate the processes of Data Boxing and Data Recording when long-term storage is required. In a similar way, users can initiate a task involving Data Scanning and Data Unboxing processes when some data stored in the long-term archival system is required to be restored.

Currently, the standard interface offered by APM is based on the open web services technology, although we are considering the implementation of other solutions.

Data Preservation in ARCHIVATOR

A very important aspect in a system like ARCHIVATOR is to be robust to software and hardware obsolescence in order to ensure data preservation for the long term. A key property of our system is that micrographic film is a true WORM (Write Once – Read Many) medium, that is, recorded data cannot be edited or manipulated. As this media is proven very stable and durable for centuries, our strategy for data preservation is not based on systematic data migrations but in the use of a permanent media and the adoption of standards for formats and interfaces, and being independent of proprietary software and of specific hardware (open technology).

At this stage of the project we are designing parts of this technology. Some of the most significant parts are:

- **Preservation operation flows**: APM currently supports operation flows specially defined for data preservation purposes. For instance, when the recording of a film reel is completed, a data object with the table of contents (ToC) of the film is created and codified in an open format (based on XML). This ToC is recorded at least twice in two different parts of the film and contains all information required to recover all recorded data objects. There exists a special operation flow that scans the whole film, recovers the ToC and reconstructs the internal database in APM.
- **Film storage format**: The storage format of data recorded on film is similar to LTFS (Linear Tape File System) [11] but specially adapted to the micrographic media. LTFS is used in LTO-5 magnetic tapes. We are currently extending our film storage format to support preservation properties, like being self-contained, self-described and extensible. For instance, there is a special control frame at the beginning of the film reel that includes all the parameters required to decode the storage format used in the rest of the film. Before this control frame, there is also space for several human-readable image frames with information useful for interpreting the film content. We are considering other solutions like SIRF [9].
- **Metadata objects**: All recorded data objects has associated a metadata object. We are developing such metadata with the preservation property in mind, taking as example solutions like PREMIS [12].

Key Innovations

The ARCHIVATOR project offers a completely new, sustainable, and cost effective alternative for long-term archiving of digital data. It is capable of archiving any type of digital data including HD audio-visual data, geospatial data, documents, and databases.

The ARCHIVATOR project uses a whole new approach to long-term archival, removing the need for constant data migration, and reducing hardware and software maintenance – both measures that create a more cost effective archival solution.

At its core, this project relies upon cutting edge innovations within the fields of data ingest, data encoding, data packaging, high speed data recording, and high precision and high speed data scanning, as well as the development of an open standard for data unboxing to effectively re-generate the original dataset.

Polyester based micrographic film is a proven archival medium and has shown extreme stability under normal storage conditions (i.e. – an archival stability of several hundred years can be expected).

All that will be required to read the information back from film will be a simple scanner, making the ACHIVATOR process entirely independent of the technology platforms used during the recording process.

User Market

The ARCHIVATOR system has the potential for broad market appeal. It will be able to be implemented into existing IT storage management systems, exist as an archival writer in its own right, or have its services outsourced via specialized service providers.

Potential users for the ARCHIVATOR system include any organization, institution, or company who consider the security and accessibility of their critically important long-term data to be a contributing factor to their future success and profitability.

Due to the special background of the partners in the ARCHIVATOR consortium, initial market introduction is planned for the film and broadcast television industry.

Leading companies in this market have internal departments to handle archiving processes. Smaller companies use service providers for archiving, offering the transfer of digital content onto the archiving material (and handling storage). Predominantly, both channels currently rely on digital storage media and the inherent migrations that system infers.

It is not only the film and television industry that have a need for long-term archiving. Almost every business, government, cultural organization, and even home user is facing the demand for long-term archival of digital data. In many segments (e.g. – governmental bodies, banks, insurance companies, medical research), such data must be kept available for the lifetime of the person or persons concerned.

The long-term archival process developed by the ARCHIVATOR project can be applied to practically any market segment with a need for long term archiving, including, but not limited to:

- National and other governmental institutions.
- Conservatories of national or international heritage.
- Libraries and documentation centers.
- Security and regulatory bodies.

- Official registers and registration offices (land and building registers).
- Insurance companies and banks.
- Industry (construction companies, energy and other supplies).
- Medical technology.
- Space science.

References

[1] "The Digital Dilemma – Strategic Issues in Archiving and Accessing Digital Motion Picture Materials," Academy of Motion Picture Arts and Sciences (2007).

[2] J. Rothenberg, "Ensuring the Longevity of Digital Documents," Scientific American, 272, 1, pg. 42-7 (1995).

[3] R.A. Lorie, "Long Term Preservation of Digital Information," Proc. ACM/IEEE Joint Conf. on Digital Libraries, pg. 346 (2001).

[4] E. Spitz, J-C. Hourcade, F. Laloë, "Lifetime of Digital Media: Is Optics the Solution?," Quantum Sensing and Nanophotonic Devices VII, Proc. SPIE, 7608, 760802 (2010).

[5] "Reference Model for an Open Archival Information System (OAIS)," CCSDS 650.0-B-1, Blue Book, http://www.ccsds.org (2002).

[6] I. S. Reed and G. Solomon, "Polynomial Codes Over Certain Finite Fields", SIAM J. on Applied Mathematics, 8, 2, pg. 300 (1960).

[7] N. Krebs, S. Rönnau, U.W. Borghoff, "Fostering the Universal Virtual Computer as Long-Term Preservation Platform", Proc. IEEE Int'l. Conf. on Engineering of Computer-Based Systems (2011).

[8] "Storage Networking Industry Association (SNIA)," http://www.snia.org

[9] "Self-contained Information Retention Format (SIRF) – Use Cases and Functional Requirements", Working Draft ver. 0.5A, SNIA LTR-TWG (2010).

[10] "ARCHIVATOR process – The solution for long-term archiving of digital data," EUREKA Eursotars project E!4683, 2009-2012.

[11] "Linear Tape File System (LTFS) Format Specification", ver. 2.0.1 (2011).

[12] "PREMIS Data Dictionary for Preservation Metadata", ver. 2.0, PREMIS Editorial Committee (2008).

Author Biography

Oscar Plata received his MSc and PhD degress in physics from the University of Santiago de Compostela, Spain, in 1985 and 1989, respectively. Since then he was associate professor in the Universities of La Coruña and Santiago de Compostela. Currently he is a full professor in the Department of Computer Architecture at the University of Malaga. His work has focused on high-performance computing and software optimization for parallel computers. Oscar is a technical advisor of Tedial.

Rune Bjerkestrand has an Engineer of Cybernetics degree and has a Bachelor of Business Administration and a Master of Management from the Norwegian School of Management. He has held senior management positions at Honeywell, Norcontrol, Kongsberg Simrad. He was Vice President of Operations and Vice President of Product and Marketing Management at Davis. In addition, he was the Director of Product Development at Infocus. Rune is one of the founders of Cinevation, where he holds the position of Director of Innovations.

©2012 Society for Imaging Science and Technology