Passive Digital Preservation
on Paper in Practice

|  |  |
| --- | --- |
| **Vincent Joguin** | **Jean-Noël Dumont** |
| *Eupalia**France**vincent.joguin@eupalia.com**https://orcid.org/0000-0003-0627-8778* | *Andra**France**Jean-Noel.Dumont@andra.fr* |
|  |  |

**Abstract – Radioactive waste management encompasses timescales ranging from centuries to hundreds of thousands of years. Therefore, it is concerned with the long-term preservation of complex analog and digital information that are key to safeguarding the environment and human health against the potential threat of radioactive waste. The multi-century timescale and criticality of the radioactive waste repositories call for information preservation strategies based on durable, robust and secure data carriers and technologies.**

**Andra, the French national radioactive waste management agency, uses permanent paper to archive analog material. Over 2020 and 2021, it tested the Micr’Olonys solution from Eupalia, a novel approach to transcribe digital data on paper for passive long-term retention and accessibility.**

**This paper presents the results of transcribing a database concerning waste packages stored at the first repository operated by Andra, producing a 464-page document instead of a million pages that plain text printing would have required.**

**Keywords – database, passive preservation, permanent paper, radioactive waste management**

**Conference Topics – Innovation; Resilience**

# Introduction

Instability is the rule over the course of centuries, with wars, natural disasters, economic crises or simply societal changes taking place usually unexpectedly but regularly. Radioactive waste management is concerned with very extensive timescales and has therefore to cope with such disruptions. Preserving information pertaining to radioactive waste and the repositories where it is stored for centuries or more significantly contributes to keeping generations to come and the environment safe from any possible adverse impact, as well as giving the opportunity to revise how the waste is to be managed over time. The richness and volume of the information concerned makes the challenge even more difficult: it can be partially preserved in analog form, but this form is unsuitable to some of it for which digital preservation needs to be considered. While mainstream preservation strategies based on migration are useful to maintain continued access to digital resources, they become simply inadequate when it comes to ensuring that material will still be accessible centuries from now, and therefore a complementary, more durable approach needs to be leveraged.

Andra, the French radioactive waste management agency, has adopted permanent paper (ISO 9706) as the reference material for long-term storage of records, after having noted that no maintenance contract with a provider of electronic solutions would be able to reach the scale of several centuries. On the contrary, permanent paper does not require maintenance and studies show that its durability in normal conservation conditions may reach at least 5 centuries. Until now, only analog contents (text, tables, pictures, diagrams) were considered for archival on paper, but this excluded more complex and larger born-digital content. Therefore, tests were conducted over 2020 and 2021 at Andra to evaluate how the Micr’Olonys solution developed by Eupalia could extend the applicability of permanent paper to digital content. Micr’Olonys was designed as a self-contained software-on-paper technology to store digital content in the form of 2D barcodes together with implementation-neutral means of access. The following sections describe the context and process of testing as it was carried out on a database relating to radioactive waste packages stored in a repository operated by Andra in the north of France.

# Preserving a Database over Centuries

## The issue of information preservation for radioactive waste repositories

Radioactive waste repositories are designed to isolate waste from the living environment without human intervention over extended periods of time. No need for human intervention does not mean that oblivion will be looked for, quite the opposite: preserving the information, data and knowledge on the repositories for as long as possible is a shared objective of the various national radioactive waste management programs [1]. This has been addressed namely by a joint initiative under the framework of the Nuclear Energy Agency (OECD/NEA) called Records Knowledge and Memory preservation across generations (RK&M), that lasted from 2011 to 2018 [2] [3], followed by a working party, Information Data and Knowledge Management (IDKM) launched in 2020 and still ongoing.

During the first centuries after closure, awareness of a repository is necessary in order to ensure that no involuntary intrusion will occur, at the time when the decay of radioactivity has not sufficiently reduced the dangerousness of the waste. Another objective of information and data preservation is to allow current and future generations the possibility to understand the repository system and its performance and make informed decisions about the repository, even long after repository closure. For example, preserving information, data and knowledge will offer future generations the possibility to review and reassess the repository performance regarding protection of Man and the environment, provide modifications or retrieve material from the repository, if they consider it is necessary, without undue costs and risks.

The RK&M initiative identified a toolbox of 35 mechanisms that may contribute to information, knowledge and awareness preservation [3], among them a dedicated set of essential records (SER), collection of the most important records for waste disposal, selected for permanent preservation during the lifetime of the repository. The SER provides sufficient information for current and future generations to ensure an adequate understanding of the repository system and its performance. This will enable responsible parties to review and verify the repository performance and the safety case, and to make informed decisions, even long after repository closure. The SER should serve as a source of detailed data and information on the repository system primarily for specialists and researchers, as well as for decision makers, regulators and other authorities. Selection criteria for the records to be part of the SER have been proposed by the RK&M expert group [4], based on a combination of relevance level (not relevant/nice to have/should have/must have) and effort it would take for a future generation to recreate the information if it were lost (some effort/extremely high effort). “Must have” information plus “Should have” associated with “extremely high effort” would be recommended for SER. Information regarding the radioactive content of waste packages is of course considered as part of the SER, but is the radioactive content of each waste package necessary? Reflections on the SER continue in the framework of IDKM, and the question of which data sets should be part of the SER is an issue that is being addressed.

## Information preservation at Andra

In France, Andra launched in 2010 the “Memory for Future Generations” program in order to address in a more ambitious and structured way the issue of long-term information, knowledge and data management for radioactive waste repositories. This concerns both the near-surface repositories operated by Andra: the Manche disposal facility (CSM), the Aube disposal facility (CSA) and the Cigéo project of deep geological repository for intermediate level long-lived waste (ILLW) and high-level waste (HLW). The CSM, situated in the northwest of France, is the first repository operated by Andra. It received waste packages from 1969 to 1994. According to the regulation, CSM is in the “dismantling and closure phase” that will still last a few decades, before entering the “surveillance phase.” The surveillance phase should last at least 300 years, allowing most of the dangerousness of the waste to disappear due to the radioactive decay. The successor of CSM, still receiving waste packages, is the CSA. Being the repository at the most advanced stage of its lifecycle, the CSM acts as a pilot for actions related to memory preservation.

For all repositories classified as nuclear facilities, the French environment code requires that a set of records will be prepared by Andra, prior to entering the surveillance phase. This set of records, similar to the set of essential records proposed by the RK&M initiative, is called “Detailed Memory File” (DMF). In order to allow for long term preservation of the DMF, the memory provisions adopted by Andra state that the DMF will be printed on permanent paper and stored in at least two different places. A preliminary version of the DMF of the CSM has been tested by a panel of external experts in 2012, providing lessons and conclusions that are presently used for its improvement [5]. The DMF contains of course a comprehensive description of the radioactive waste inventory, but at this stage not the database for each individual waste package, because printing this database on permanent paper would require an excessive amount of paper and of storage volume in the archives. The data are kept electronically, which is also the most useful for present needs; they will be migrated regularly, for at least as long as Andra will exist and manage repositories, which is expected to last still for more than a century.

The “Memory for Future Generations” program [6] aims both at implementing robust memory preservation provisions, based on the regulatory requirements but not limited to what is explicitly required, and leading R&D activities to extend the robustness and timescale of memory preservation. Apart from archives-related activities, it has developed a wide range of societal interactions, including for example artist contests and residencies, in order to spread in the society as largely as possible the awareness of the repositories and related documents. R&D activities are multidisciplinary, dealing with social sciences and humanities (e.g. semiotics or socio-anthropology) as well as landscape archaeology or materials sciences. Not to forget reflections on long-term preservation of digital archives, including the tests of database transcription on permanent paper presented in this publication.

# Preliminary Test

Preserving the database mentioned above for each individual waste package stored at the CSM repository in analog form would translate into a printed document of about one million pages, a volume too large to be practical for long term archiving, but also for efficient access to relevant information. Therefore, Andra considered the Micr’Olonys solution developed by Eupalia to transcribe this database into digital form on paper.

Micr’Olonys prints a simple self-contained primer together with 2D barcodes (called “emblems”) that contain a digital file. The Micr’Olonys primer gives future users the ability to restore the information without needing any specific technology such as particular hardware, operating system or programming language.

A general overview of the passive digital preservation approach and Micr’Olonys is provided in [7], while the Micr’Olonys technology is introduced in more detail in [8].

A preliminary test of Micr’Olonys was conducted at Andra in September 2020. A subset, consisting of 4,042 lines of data related to CSM radioactive waste was exported from the complete database that contains around 1,5 million lines into a Microsoft Excel XLSX file that was 1.7 MB in size.

For this test, the Excel file was then converted to the FODS format: Flat OpenDocument Spreadsheet. This uncompressed, fully textual format is supported by the LibreOffice suite. Its very explicit XML structure would be an advantage to make readability and reinterpretation easy over the long term. The test database subset produced a 41.8 MB FODS file.

This file was then compressed using a prototype version of the integrated Micr’Olonys compression algorithm to strip its size down to 111 kB (a 99.7% downsize). The very high compression ratio could be reached because of the strong redundancy of both the FODS format and the data the file contained. This file was transcribed into 3 emblems over 3 A4 pages: 2 emblems for data, and 1 redundancy emblem for critical error correction such as one missing page.

The French advanced prototype version of the Micr’Olonys primer was appended to these emblems to form a complete self-contained 15-page paper document. The primer consisted of about 3 pages written in plain French, 4 pages of pseudo-code written in simple French algorithmic language, a list of encoded letters over 2.5 pages, and 2 system emblems over 2 pages.

The 2 system emblems contain, in compressed form, all necessary code to parse and decode the emblems that contain user data, including the appropriate decompression and error correction functionalities. This code runs within the computing environment that implementation of the pseudo-code creates.

The 15-page document was handed over to a computer engineer at Andra with absolutely no prior knowledge of Micr’Olonys or what this document was about. The engineer was tasked with “making something” out of this document, by supervising a 2nd year co-op student working at Andra that would take care of actual implementation.

Together, they managed to fully restore the original FODS database file in about two weeks of discovery and software development. In the process, they faced a few minor challenges that they eventually solved autonomously within a few days.

Two main shortcomings caused these challenges that were unforeseen by the Micr’Olonys authors. One is that recent changes to the Micr’Olonys primer structure created uncertainty as to how the software was meant to behave at some point once implemented, and gave the impression to the testers that their implementation was incorrect, when it was actually fully correct. The other is that the end condition and guidance to make use of the restored data was insufficiently explained, again giving the testers the impression that “something more” was to be expected, when actually they had reached successful completion of the decoding and restoration process.

Both shortcomings – and some other minor ones – were addressed with a subsequent update to the Micr’Olonys primer. It is expected that this improved update would accelerate implementation for other testers, shortening the necessary time to about one week for a junior programmer – and probably also for a senior programmer in the distant future.

This test demonstrated the relevance of storing digital information in such documents that make it possible for a junior software developer to restore the original data bit-for-bit without needing any specific hardware or software. Indeed, such a solution, in addition to being independent from today’s technology, should also not assume current advanced know-how will remain readily available among professionals of centuries to come. Submitting the document to a beginner in programming, who has not yet acquired today’s best practice and conventions, is to our mind the best way to come close to what the process of implementation by a future programmer is likely to be.

With the restoration process validated on a small test case, implementation of the Micr’Olonys solution at Andra continued in 2021 with the transcription of the complete database.

# End-to-end Database Transcription to Paper

The complete database consists of a main table containing around 1.5 million lines, and a few other much smaller tables containing complementary information and metadata. It was decided to transcribe each table individually so that each produced document would remain simple in its structure. Focus was on the main table as its size allowed for convincing validation of the approach, should it succeed.

The question of the stored file format was of course important. Making sure the format of the restored data will not hinder its reinterpretation, either manually or automatically, is a key element in the more general requirement of ensuring the whole document remains meaningful over time, without any modification or update to it whatsoever. The native Oracle format of the database was not an option as it is complex and proprietary. The richness offered by the FODS format selected for the preliminary test, and other XML-based formats, didn’t seem relevant for a single flat database table. Therefore, the CSV format (comma-separated values) was selected for its simplicity and relevance. Moreover, it is a widely used and supported format within the database community, and the French National Archives routinely use this format to archive databases in their custody.

Extraction of the main database table into the CSV format generated a 626 MB file. The complete compression and decompression feature was implemented into the Micr’Olonys solution, both in the transcriber software and in the primer, the latter incurring absolutely no additional complexity. The compression feature is fully transparent to the user, both when transcribing data today and when restoring it in the future.

Using Micr’Olonys built-in compression, the 626 MB CSV file was brought down to 23 MB, a 96.3% reduction. This reduction was in line with expectations following the compression ratio of the preliminary test. This compressed data was then transcribed into 444 emblems in as many A4 pages, including redundancy emblems (3 emblems in addition to every group of 17 emblems).

Regarding the primer, it was decided that a bilingual document, written in both French and British English, would improve chances of its meaning coming out clearly in the distant future, when the form and practice of natural languages will have changed to an extent that is likely to impede comprehension. Indeed, it is widely known that the Rosetta stone which presents the same message in two different languages, Ancient Greek and Ancient Egyptian, the latter with two different transcriptions, hieroglyphic and Demotic characters, was key for Champollion to deciphering the Egyptian hieroglyphic writing.

The primer was structured so that both language versions integrate seamlessly, referencing appropriate page numbers and common sections (the list of encoded letters and the system emblems).

The bilingual primer being 20 pages in length, the complete document produced for the main database table amounted to 464 pages, a very satisfying volume when compared with the estimated one million pages of the same information written in human-readable form.

In addition, the “emblematic” form offers other advantages compared to the analog form. For one, built-in error correction can easily make for small physical alterations to pages, while the same alterations to the analog form would mean irremediable loss of information, unless the document is replicated. Moreover, restoration from the emblems immediately producing information in digital form makes it possible to automatically process the data, in a trusted manner since the restoration software built in the primer will indicate the presence of even the slightest error of one byte, or its absence thereof. A comprehensive search functionality is also made possible very easily for a human user to quickly find relevant information within the mass of data. On the contrary, an analog form would translate into a burdensome and error prone process of Optical Character Recognition, or the necessity to manually browse through a million pages in search for the useful information.

Printing the document of 464 pages was carried out on a standard professional Toshiba multi-function laser copier. Micr’Olonys prints each page at the most common 600 dpi resolution, each emblem being formed strictly from black and white elements. From this document, the 444 data emblems were scanned using the automatic document feeder (ADF) of the same multi-function copier. Scanning was done at 600 dpi as bitonal (black and white without greyscale), again a very common and widespread configuration. Using the ADF means quick scanning, but also some geometric distortions as the pages are moved imperfectly for scanning by the mechanical device. Fig. 3 & 4 below show such distortions.



Figure 3 A squeezed line, visible in the middle, resulting from imperfect scanning. This emblem is still read without any error.



Figure 4 A major distortion resulting from imperfect scanning.

Using the integrated Micr’Olonys reader software provided by Eupalia, a first decoding pass resulted in 17 of the scanned emblems with errors. The corresponding pages were rescanned but 4 remained in error. However, since they were distributed within different groups of 20 emblems, complete error correction could have been achieved, and a bit perfect file could have been restored at this stage. To complete the testing process, the remaining emblems were rescanned using the flatbed scanner, making sure the emblem images would not be cut at the edges, and using greyscale instead of bitonal. As a result, all emblems were finally correctly decoded by the Micr’Olonys software, and the stored file was restored. A file comparison with the initial file validated that the process was successful.

Using the primer algorithm implementation in C# that is integrated within the Micr’Olonys reader software, restoring the initial file took 3 h 10 min, for an average of 55 kB / s. This was executed on a laptop computer equipped with an Intel Core i5-6300U processor running at 2.5 GHz.

# A Complete Technical Chain to Ensure Durability

A passive digital preservation solution needs to ensure that each and every aspect of it will stand the test of time, without any human intervention. It is a chain where any weak link endangers the whole process.

For the experiment described here, we used permanent paper, a medium that is able to last at least 500 years in normal conditions of conservation. Film is another medium with a similar life expectancy.

The Micr’Olonys primer is a short document that only makes use of fairly simple and widespread concepts of computer science and document imaging that have been in use for decades. The description parts in natural language have been written using common words rather than specific technical words, so as to mitigate the risk of meaning shifting over centuries.

The CSV format was selected to store the archived information, CSV being a textual format with a very simple and straightforward structure that can be easily described.

Finally, an opening part puts the document in context. A relevant UTF-8 table (i.e., one only presenting those characters actually present in the stored document) is the key to converting the digital information into information understandable by humans, while the CSV format specification is the key to making the data structure explicit.

Of course, the stored data should be put in context: what does it describe or represent, how should it be used and in which context, is it related to other documents or information, etc.

# Conclusion

Storing digital information in a passive way, one that does not require regular human intervention and physical refresh, is mandatory when considering multi-century timescales. The radioactive waste management sector is concerned with such timescales (and more extended ones) and is mandated to preserve and transmit critical information over to future generations.

In this paper, we presented how Andra, the French radioactive waste management agency, worked together with the Eupalia company to test and validate Micr’Olonys, a passive digital preservation solution that extends the applicability of permanent paper to digital content.

During the course of 2022, the collaboration will continue with the testing and validation of Micr’Olonys on microfilm. Indeed, information density significantly improves with microforms when compared to paper. Moreover, the completely different physical natures of film and paper has the potential to open up a more robust passive preservation strategy by using both media in combination.

Applying Micr’Olonys to other emerging passive digital preservation media, such as ceramic and DNA, is also being actively pursued. In particular, these media have the potential to offer greater longevity and density over paper and film, and would further diversify digital preservation media to increase overall durability and accessibility, thus improving resilience of the approach.

#### 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 was supported by the technical industries financial instrument of the Centre national du cinéma et de l’image animée (CNC).

#### REFERENCES

1. J.-N. Dumont, S. Wisbey, S. Hotzel, A. Berckmans, and A. Claudel, “Analysis of the Needs for Long Term Memory and Knowledge Preservation Relating to Radioactive Waste Disposal Facilities,” *WM2017 Conference*, March 5 – 9, 2017, Phoenix, Arizona, USA.
2. S. Wisbey, et al., ”Development of an Integrated System for the Long-term Preservation of Records, Knowledge and Memory relating to Radioactive Waste Disposal Facilities,” *WM2017 Conference*, March 5 – 9, 2017, Phoenix, Arizona, USA.
3. Radioactive Waste Management and Decommissioning, “Preservation of Records, Knowledge and Memory Across Generations: Final Report,” OECD/NEA, 2019.

https://www.oecd-nea.org/jcms/pl\_15088

1. Radioactive Waste Management and Decommissioning, “Preservation of Records, Knowledge and Memory (RK&M) Across Generations: Compiling a Set of Essential Records for a Radioactive Waste Repository,” OECD/NEA, 2019.

https://www.oecd-nea.org/jcms/pl\_15090

1. J.-N. Dumont and G. Martin, “The First International Decennial Appraisal of the Detailed Memory of the Manche Disposal Facility,” *WM2014 Conference*, March 2 – 6, 2014, Phoenix, Arizona, USA.
2. J.-N. Dumont, P. Charton, and F. Boissier, “Andra’s Strategy and Approach for Long-Term Memory Preservation of a Deep Geological Repository,” *WM2015 Conference*, March 15 – 19, 2015, Phoenix, Arizona, USA.
3. V. Joguin, “Passive Digital Preservation Now & Later: Microfilm, Micr’Olonys and DNA,” *iPRES 2019*, Sep 2019, Amsterdam, The Netherlands.

https://ipres2019.org/static/pdf/iPres2019\_paper\_139.pdf

1. R. Appuswamy and V. Joguin, “Universal Layout Emulation for Long-Term Database Archival,” *CIDR 2021*, Jan 2021, Chaminade, USA.

http://cidrdb.org/cidr2021/papers/cidr2021\_paper30.pdf