
DIGITAL PRESERVATION TECHNIQUES TO FACILITATE KNOWLEDGE MANAGEMENT IN THE NUCLEAR SECTOR

M. Claxton, R. Sharpe

Tessella Support Services plc., United Kingdom

E-mail address of main author: Mark.Claxton@tessella.com

Abstract: The advances of computer technology offer many new opportunities but also new challenges. One of the most crucial challenges is tackling the problem of digital preservation: ensuring that the digital information we create and store today will continue to be accessible for as long as we may need it. Industry commentators have raised the prospect of a 'digital dark age' stretching from the late 20th to the early 21st century, as huge amounts of digital information are at risk of being lost.

The problem is becoming well known:

- Storage media such as magnetic tapes and disks and CD-ROMs have a finite lifetime.
- Storage media formats change rapidly (consider how few people still have a 5.25 inch floppy disk drive).
- Most challenging of all, the file formats in which information is stored rapidly become obsolete as new software packages appear and older applications become unsupported and unavailable.

In other words, if the lifetime of digital records exceeds that of any part of the software/hardware stack used to create them then the owners of those records face a potential digital preservation problem. Tessella has been working with government and private sector organizations to develop solutions to the problems of digital preservation.

This presentation outlines the problems and potential solutions based on Tessella's experience working with the UK National Archives, the Dutch Nationaal Archief, and most recently the US National Archives and Records Administration (NARA) on systems and strategies to ensure that 'born-digital' records are stored and managed safely and securely, and continue to be accessible into the distant future. This is essential both to maintain the accountability of governments, for our cultural heritage and to meet regulatory compliance in areas as diverse as the Nuclear, Aerospace and Pharmaceutical sectors. Tessella are actively researching innovative strategies to ensure that digital information can survive reliably as hardware and software technologies come and go. The paper uses examples of platform migration strategies used to ensure reliable access to nuclear research data as well as data management strategies for the long term acquisition, management and access to geosphere data used in the selection and design of waste repositories.

1. The Problem

The Nuclear Industry presents unique challenges for information storage and retrieval. Data contained in documents and files created decades ago, as well as those created today, must be available throughout the lifetime of the Nuclear facility, and potentially remain usable for hundreds if not thousands of years. The ability to retrieve this data is critical for:

1. Ensuring the continued safe operation of existing plant while supporting the decommissioning process.

2. Providing a reliable source of information for the wide range of regulatory and commercial organizations involved in the Nuclear sector
3. Allowing a knowledge pool to be made accessible to future generations of Nuclear engineers and regulatory bodies

Ideally, information would be stored in a manner closely integrated with current site operations or anticipated decommissioning processes, in a format that can be accessed easily far into the future. However, no single technology currently provides this capability.

Whilst this kind of long-term data storage may be new to the nuclear sector, groups as diverse as pharmaceuticals, aerospace companies and governments are already benefiting from digital preservation. What lessons can be learnt from other projects in this field?

Paper records, like photographic solutions, have a place in the long-term archive strategy, but both can degrade over time. More importantly, they lack the flexibility of digital storage – the ability to search, mine and link information together facilitating Knowledge Management. Digital records can hold a wide variety of information unsuited to paper record keeping – 3D CAD models of installations and material repositories, databases of personnel exposure records, audio recordings of meetings, or datasets holding experiment results. Over the past decade the Nuclear sector has embraced the power of digital information storage, and so any archiving solution needs to be built on these digital records from the ground up.

Given the importance of this digital data, how can you be sure that it is being saved and preserved with long-term, protected access in mind? Today's systems may not be able to search or even read old data. Tomorrow's systems may not be able to access today's. The impending 'digital dark age' has implications for everyone in the Nuclear sector.

One issue is the nature of electronic storage media itself but this is relatively straightforward to solve (for example by transferring records on degrading electromagnetic media to fresh media). The real issues are those surrounding data formats, and the ability to retrieve accurate information in the future when the software used to create it may no longer exist.

Critically, digital preservation and archiving is more than just record management. It is not sufficient to preserve the bitstream, but rather the content. Research work has been carried out over recent years to identify the best strategies for long-term retention of records. Tessella's involvement with the Dutch Government's 'Testbed' research project[1], and our design work on digital archives for other governments and organizations, has provided a set of principles for implementing these solutions, which we present below.

As the lifetime of data is measured in centuries, we can be sure that the software used to originally generate and access critical records will no longer be current by the end of the retention period. If an organization is unable to access records because software or hardware no longer exists to read them, the consequences may be severe.

2. How to Resolve the Problem

What does digital preservation mean in practice? Based upon the authors' experience of long-term digital archiving, the following are all essential components of a digital preservation strategy.

2.1. Set default retention schedules

Every functional class of record requires a default retention schedule, and this applies for both the original documents and the underlying data that has been extracted and retained. For example, setting a schedule of 'maintaining on-line storage during the lifetime of the

decommissioning project, then move to off-line storage for 200 years' ensures that commitments are made up-front to the long-term preservation of data. Records required by regulators in support of safety will need to be accessible for the lifetime of the decommissioning process, so this entails a commitment to retain these records for at least the scheduled term even if the technology used to produce and access them becomes obsolete.

This long-term retention can be under-developed and under-funded in the commercial sector, exposing companies to the risk of long-term failure due to inadequately stored records. Early investment and review in a digital archiving program is required to mitigate this risk, and it is likely that public accountability requirements will demand evidence of the reliability of electronic record stores.

There will also be occasions when particular records need to override this default retention scheme – for example, when new research alters the originally conceived approach to decommissioning. Flexibility should be designed in from the ground up.

2.2. *Maintain electronic files*

Electronic records are currently held on magnetic or optical storage media such as hard disks, tapes or optical disks. Such media can be subject to degradation in ways that paper records are not. Ink may fade over time, but a tape can lose data due to magnetic fields from the next layer of the spool, or even by physical degradation if not properly stored. Studies from media manufacturers indicate typical lifetimes for these media, and so can be used as a basis for transfer schedules.

Media maintenance should be a continuous process of management, combined with planned integrity and migration testing. Once again, flexibility should be designed in to cope with unexpected problems. Original compact discs sold in the 1980's were marketed as 'keeping your music safe forever', but already they are showing previously unpredictable signs of decay. A well-managed system should adapt to these changes seamlessly.

2.3. *Maintain content*

The Dutch Government's 'Testbed' project, involving Tessella specialists, investigated approaches to long-term maintenance of the ability to access electronic records. With increasingly short production cycles in software, file formats that are ubiquitous today may become obsolete in only a few years – far shorter than the retention periods required by nuclear sector. A record that cannot be accessed is useless, and the data within is effectively lost.

Lessons for the nuclear industry have also been learnt from Tessella's work with the UK National Archives, notably the need to have a flexible framework for separating *conceptual content* from a particular file format or storage mechanism. This approach isolates underlying data from the natural obsolescence of any particular technology. Filesystem hierarchies should be seen as artefacts of current technologies and thus subject to change over time as new technologies evolve. This approach removes the assumption of permanence of any particular technology, and facilitates simpler migration strategies, for example moving a Paradox database to Oracle.

Records can hence be supported in a number of technologies, with a well-defined policy determining the number of concurrently supported formats. Migration allows records to be kept in a structured framework that can adapt to new technologies as they arrive. Whether proprietary formats or common open standards are used, migration along planned, minimal-change pathways ensures that the underlying conceptual content remains as accurate a representation of the original as technological change will allow.

2.4. Drive migration by policy

Advanced planning of digital preservation should be dynamic, updating itself as technology evolves alongside the project. Flexibility is key – the future is uncertain, particularly in terms of the technology that we will be using. Who could have predicted the widespread use of computers we have today fifty years ago, or the Internet twenty years ago? Data formats are just as prone to change. Some things we can predict; we can expect that more and more migration technology will become available, as increasingly large amounts of digital information are produced. Some of this change will be driven by the nuclear industry, with its extreme digital archiving requirements.

As migration technology becomes more advanced, we can migrate more records and keep them active and accessible. We can re-migrate older records where newer migration methods allow more faithful reproductions of original documents. To harness this power and provide a robust framework for continual digital preservation, it is advisable to have a coherent ‘policy engine’ that can suggest migration pathways with minimal changes to formats that are approaching obsolescence.

Tessella’s work with the UK National Archives included the development of the PRONOM file format database, one such policy engine [2]. By storing information on file formats themselves (including predicted obsolescence periods), migration pathways and the tools needed to read these formats, PRONOM can generate and support a variety of migration strategies. This effort complements the Global Digital Format Registry (GDFR)[4], currently at the conceptual stage and hosted at Harvard. Tools such as these can be vital for judging the benefits of migration against the potential loss or alteration of data that migration may introduce.

3. An IT Based Solution to the Problem

The Nuclear industry can learn from experience in other sectors. The long-term preservation of digital material will be critical to the continued safe and efficient operation of Nuclear facilities and decommissioning programmes. This requires a structured approach to the retention of records, and the ability to access them day-to-day in a way that preserves content and meaning

3.1. Open Archival Information System: a solution framework

In order for different organizations to share digital preservation experiences and learn from each other, it is essential that each solution can be compared. However, digital archiving is a relatively young discipline and, as such, standards are in their infancy. Nonetheless, ISO are encouraging the development of good practices and have endorsed NASA’s Reference model for an Open Archival Information System (OAIS) [3].

OAIS splits the problems of archiving into six functional entities as shown in figure 1. These are:

1. **Ingest.** This covers the issue of getting records into an archive, including the capture of appropriate metadata to allow them to be found, extracted and meaningfully used in many years’ time
2. **Data Management.** This covers the controlled editing of data input into the system
3. **Storage.** This covers the issue of physically storing records in an archive, including the creation of an appropriate backup policy, regular media migration etc.

4. Access. This covers two related aspects: finding records within the archive and disseminating them to consumers. This includes ensuring that the appropriate information is only disclosed to appropriate users of the system
5. Preservation Planning. This involves ensuring that the contents of an archive remain more than just a meaningless bit-stream
6. Administration. This covers the running of the system itself including its maintenance

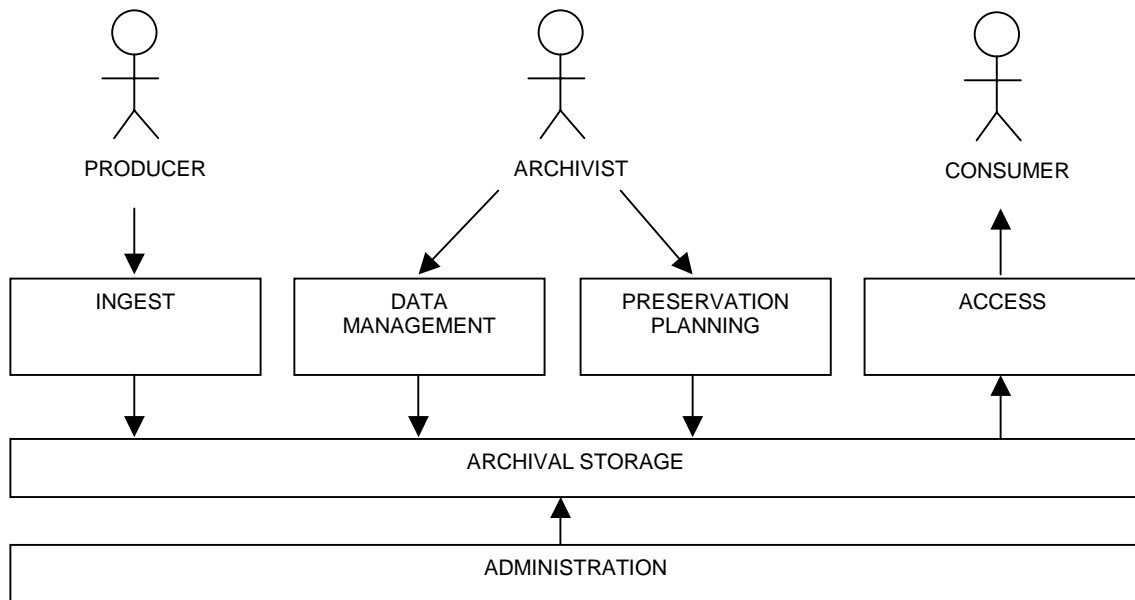


Figure 1: Schematic view of the OAIS model

Of the 6 steps described above, Preservation Planning is arguably the most important, with 3 viable options open to the Nuclear Industry. All solutions have one fundamental thing in common: the original files in a record are always maintained for as long as the record needs to be retained.

3.1.1. The museum approach

One possibility would be to maintain the old hardware and software used to create the data in the first place. However, this is not very practical. Such a solution would require the maintenance of every combination of hardware and software required, the hardware would become increasingly expensive to upkeep and would eventually become irreparable. This is really only an interim measure.

3.1.2. The migration approach

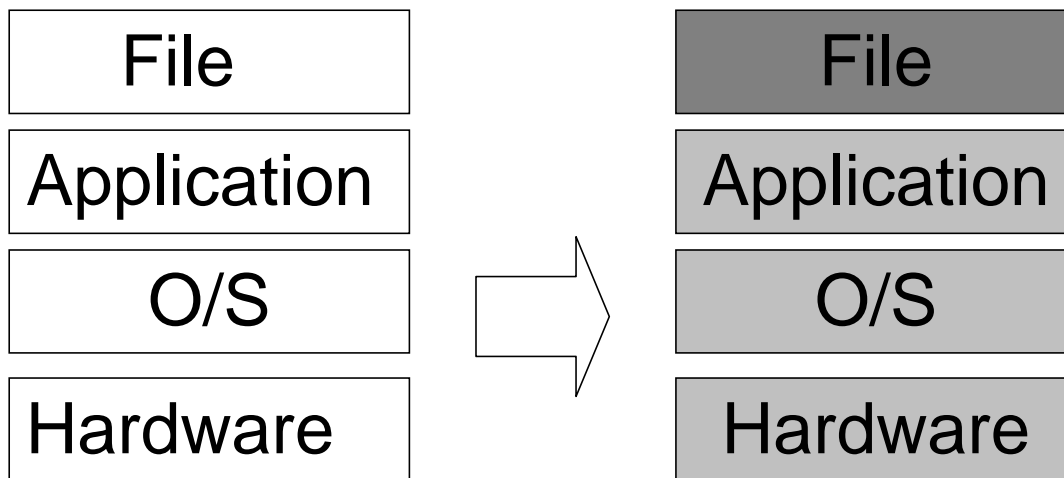


Figure 2: Migration involves accepting that natural changes occur to hardware, operating systems and application software (light grey changes) and therefore the original file is deliberately transformed (dark grey change) in order to allow a record to remain readable

In this technique, see figure 2, a copy of the original is transformed into another, more modern format that can be read by newer application software (potentially running on updated hardware with a newer operating system). For instance, scientific data in a bespoke binary format may be transformed into a document conforming to an XML schema, which (as it is self-describing and based on very simple low-level technology, i.e. it is fundamentally just a simple text file) is less vulnerable to obsolescence. In more complex cases it may be necessary to perform a series of transformations over the lifetime of the data, either because of a change in the available application software or because a better transformation engine becomes available. In such cases, it is normally preferable to return to the original file and transform this into the new format, rather than transform it from the previous migration (as this will potentially already have lost some of the information in the original).

The fact that migration involves a transformation which may result in loss means that it is necessary to understand and categorize this loss so that different transformation software can be assessed and compared. The attributes that need to be considered can be split into five categories:

1. Context. This is set by metadata and thus is unaffected by migration (although the migration process should itself be documented).
2. Content. A good transformation should preserve all the content of the original. However, sometimes the new format will not allow information to be kept in exactly the same form.
3. Structure. It is important to remember that, if an accession undergoes migration, for either preservation or presentation purposes, the logical (technology-independent) structure will be preserved, but the physical (technology-dependent) structure may be altered as not all migrations will lead to an exact 1-to-1 file correspondence. This means that migration is potentially a complex process and as such could be prone to human error (e.g. marking a file incorrectly as having been superseded by a newer version).
4. Appearance. It is quite hard to preserve the look and feel of the original when performing a migration. For most purposes, this may not matter too much but there is not always a

clear-cut distinction between appearance and content. For instance, if an author uses bold or italics at some point in a document, it is probably an emphasis and thus can be interpreted as being part of the content of that document.

5. Behavior. One of the advantages of digital records is that it is possible to manipulate the information within them. For example, database records can be queried to provide new views of the information contained within them or a model can be re-run using different initial parameters. This aspect of a digital record relies on programming logic embodied in the application software and is thus difficult to preserve by migration.

One of the key aspects of preservation planning is ensuring that the strategy for data types is reviewed regularly (e.g. a strategy for a given data type that is relying on the use of a given piece of application software will need to be reviewed if support for that application ceases). This means that there is a requirement to maintain a repository of information about each file format stored in the archive, to assist archivists in determining its best preservation strategy (e.g. to plan when each format will need to be migrated). This strategy may evolve with time as better technologies become available. With assistance from Tessella, the UK National Archives has created such a library (called PRONOM), designed to share information with other archiving organizations and to allow anyone to submit information on new formats [2].

3.1.3. The emulation approach

This technique potentially has an advantage over migration in that it should allow the look and feel and behavior of the original application to remain intact (whereas these are potentially lost in migration). This will be especially helpful for records with a high degree of behavioral content (e.g. virtual reality models). Also, for a given piece of hardware, such an emulator can be written once and re-used by many organizations (although an emulator may need to be re-written when hardware changes again). However, such generic emulators do not yet exist, thus the concept cannot yet be seen to be a proven universal approach. The approach also means that licensed copies of the original application software (in fact a record may rely on lots of applications to operate as originally intended) and the original operating system must be retained, including the relevant bug fixes, service releases etc. It also means that the effort required to access an old record could be considerable, since the original application software and operating system must be installed together with the emulator before the record can be meaningfully interpreted.

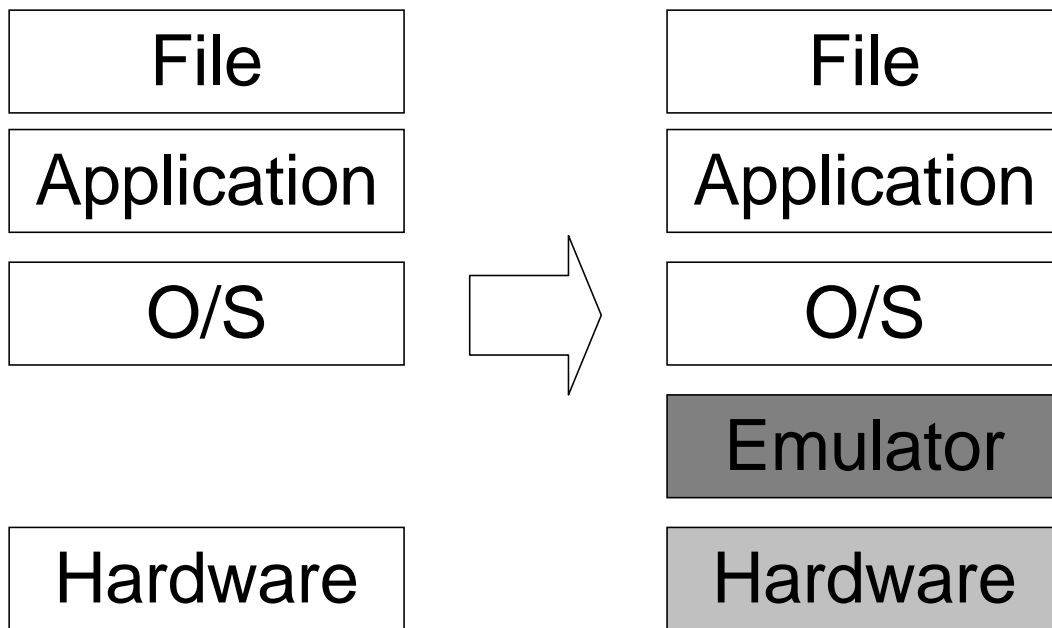


Figure 3: Hardware emulation involves accepting that natural changes occur to hardware but makes no change to the original file, application software or operating system. Enabling the software to continue to run requires the creation of an emulator (dark grey change) to emulate the original hardware on the new hardware

Figure 3 demonstrates Hardware emulation, however there are alternative emulation methods:

- **Operating system emulation.** In this case, we retain the original file and application software and accept ‘natural’ evolution of both the operating system and the hardware
- **Application software emulation.** In this case, we retain the original file and accept ‘natural’ evolution of the application software, operating system and the hardware

Based on research undertaken by Tessella, neither of these currently seem as feasible as hardware emulation (see the summary of the result of the Dutch Government Digital Preservation Testbed project for more detail[1]).

3.2. Benefits of Open Archival Information System

The specific benefits of the framework outlined above include:

- Storing the original, unmigrated record maintains the ability to produce original records in accordance with public accountability responsibilities, as well as allowing better migration later when technology develops
- Storing the current version, in contemporary formats, makes knowledge available for data mining and leveraging in new research as time moves on
- Flexible indexing policies allow cross-format searching to find useful data and documents
- Driving the ongoing preservation by policy, based on the current state of technology at any given point in the retention period, ensures that the system remains flexible and meets the business needs at any given point. Balancing migration risks against business benefit guarantees the system is always working in the organization’s best interests

Tessella's experiences in archiving across industries [5] show these to be principles that apply across domains, applying as much to record retention strategies for the nuclear industry or drug development as for government and corporate records.

REFERENCES

- [1] DUTCH ARCHIVES 'Testbed' research project website
www.digitaleduurzaamheid.nl.
- [2] THE UK National Archives website www.nationalarchives.gov.uk/pronom.
- [3] OAIS website <http://ssdoo.gsfc.nasa.gov/nost/isoas/overview.html>
- [4] HARVARD University Library GDFR website <http://hul.harvard.edu/gdfr/>.
- [5] TESSELLA website www.tessella.com/Services/Discipline/digital_preservation.htm