



DIGITAL PRESERVATION: PRACTICAL EXPERIENCES

Robert Sharpe

TESSELLA SUPPORT SERVICES PLC

Issue V3.R1.M0

January 2005



1. Executive Summary

The preservation of digital material is an issue for any organization which produces electronic records that have a lifetime exceeding the lifetime of the software used to create or read them. The challenges involved are considerable since the aim is to be able to read any electronic record in the future regardless of the original format, software application(s), operating systems and hardware used to create it. Failure to address this issue could lead to irrevocable loss of the intellectual property of that organization.

In recent years, a number of mainly academic and government organizations have been at the forefront of facing these challenges, which has allowed a body of knowledge and experience to be developed that can give other organizations the confidence to address the issue.

Tessella has played a key role in a number of the most practical of these initiatives. As a result, we believe that it is possible to take steps to safeguard the digital knowledge of an organization and thus preserve digital records using today's technology. We recognize that such a solution will not be complete and some further experimentation and research is required, but we believe that this can occur in parallel with active preservation measures. The alternative would be to wait, but the greater the time lag between creation and facing the issues involved, the greater the risk of information loss or excessive expense in order to be able to retrieve such information. Taking proactive measures also allows organizations to gain invaluable experience of the processes involved whilst evolving a more complete solution to their digital preservation problems.

This philosophy leads to us recommending a solution consisting of modular parts based around a framework. The framework gives the required structure to product and process, whilst the modularity allows third-party solutions to be added as time goes by, thereby leveraging digital preservation solutions achieved elsewhere.

2. Introduction

2.1 What's the problem?

In today's society an increasing amount of information is being created and stored digitally. In addition, the number of ways in which this information can be stored is increasing, with new software products or new versions of existing products constantly coming onto the market. All of this means that there is a large quantity of digital records, potentially containing information of vital corporate or societal importance, in an ever expanding number of formats.

However, these formats rapidly become obsolete, and as hardware and operating systems move on, digital files can become unreadable. In many cases the format can be read by newer software but even then some of the information in that file may be altered or lost in the transformation. This means that most organizations' own information is in danger of becoming inaccessible only a relatively short time after it was created.

Digital records can also be endangered because of the way in which they are stored: the media can deteriorate or become difficult to read owing to the obsolescence of the associated hardware. This is in sharp contrast to paper records which, provided they are stored in the appropriate conditions, are likely to remain readable for a very long time.

A case study illustrates the problem well. The Domesday Book, William the Conqueror's survey of England in 1086, is still readable today in the UK National Archives at Kew, London. A modern, digital version of the 'Domesday Book' was created by the BBC children's program Blue Peter to celebrate the 900th anniversary of the original. Children were invited to submit digital material about their community, which was stored on the latest technology to guard against obsolescence: 12" laser disks. Just fifteen years later, serious action had to be taken to save these records from being lost for a number of reasons.

Firstly, the media has become very difficult to read with relatively few laser disk readers still in working order. Fortunately, the time-gap from creation was still sufficiently small that some readers still existed, so this problem could be solved (indeed the National Archives had a working system). This allowed the files to be extracted in binary format onto more modern media. The next challenge was to interpret this so that it was not just a meaningless string of 1s and 0s, as the digital files could not be interpreted by any modern software.

Solving this problem was a not a trivial exercise but was completed successfully after considerable effort, including experience and input from the original record creators.

For more details on this task, see www.nationalarchives.gov.uk/preservation/research/domesday.htm/default.htm.

Although the records are now safe, for another 15 years maybe, the amount of effort required to preserve a relatively small amount of information shows it is not practical to rely on such methods for all digital records.

2.2 Does it affect me?

The question every organization needs to ask itself is whether the usefulness of its digital material will outlast the lifetime of the software used to create or read it. Such application software will rely on an operating system and hardware, and this typically has a lifetime of just a few years. Thus, the answer to this question is almost certainly: yes!

All organizations are threatened with the loss of information, which will be very expensive to recreate, but it is mainly those that are required to save data for a long time that have been at the forefront of addressing the challenges of digital preservation. These organizations include:

- Scientific organizations that collect large amounts of data over a long period of time. For instance, if a mission is sent to collect data from a remote planet, it is definitely worth going to the expense of preserving the data. One of the advantages that scientific organizations have is that they can often dictate their own data formats
- Universities and academic research organizations. A lot of money is spent generating and gathering research information, and such institutions have a duty to keep it for future generations of researchers. For example, a number of university libraries are creating digital collections of thesis abstracts, and in some cases are saving full theses (again, aided by the fact that they are able to dictate the format)
- Regional, national and international libraries responsible for maintaining collections of new materials: almost all of which are now produced digitally. These organizations have a difficult job although most of the materials they receive will be from a relatively few sources (e.g. the publishing industry) and will be in well-defined formats

- Regional, national and international archives responsible for maintaining government records. Like libraries, these organizations face the issue that almost all of their future ingest material is now being produced digitally. In this case there is also likely to be a wide variety of formats with little uniformity. Some standardization may be possible, but this can never be 100% effective since it is likely that there will be a requirement to archive information that has originated from outside the influence of these standards
- Companies in regulated industries, notably in the pharmaceutical and nuclear sectors. These industries have a duty to preserve information often in a wide variety of formats, some of which (such as data output from a given piece of equipment) rapidly become obsolete

2.3 What can I do about it?

The result of the research and active preservation measures performed by the types of organizations described above means that it is now possible to plot a roadmap, which other organizations can follow to solve the digital preservation problem.

Not all parts of this roadmap are well defined and, even in the better defined areas, it is expected that the tools needed to perform the job will need to continue to evolve to meet the challenges of preserving information stored in tomorrow's formats. However, it is possible to take active measures today to preserve current digital content. This is important since, as the BBC Domesday project shows, the longer the problem is put off, the greater the risk of losing information or incurring excessive expense in an attempt to preserve material.

This white paper outlines a digital preservation solution, including specific knowledge gained by Tessella from working on a number of archival projects as outlined in section 10. The solution is based upon the Open Archival Information System (OAIS) framework (described in section 3) and work performed by the organizations mentioned above.

3. 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) (see <http://ssdoo.gsfc.nasa.gov/nost/isoas/overview.html>).

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

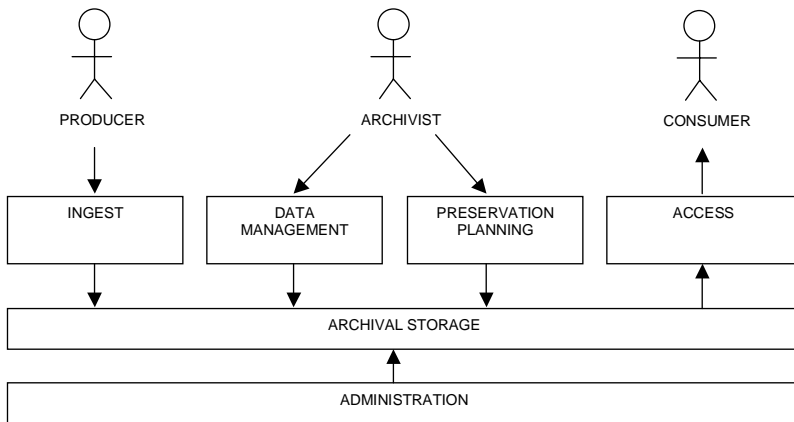


Figure 1: Schematic view of the OAIS model

Tessella has experience of building practical archival solutions that cover each of these functional entities (see section 10). In particular, Tessella has built both the UK Digital Archive and a file format repository (called PRONOM) to the UK National Archives' specification after open tenders. The rest of this document discusses the lessons learned as a result of this experience.

4. Ingest

For a digital archive to be effective, material must be ingested into the archive. This requires material to be selected, stored in an appropriate structure, and described by appropriate metadata.

4.1 Selection of records

Before archiving a set of potential records, it is necessary to decide whether they are worth keeping. One option, of course, is to archive everything. However, the cost of archiving is large: archiving a file is a commitment to keep it for an exceedingly long time (potentially in perpetuity) in a managed, well-maintained environment. In some cases it may not be possible to archive everything for security, intellectual property, legal or other reasons.

The ideal situation would be to review all material, decide what needs to be archived and reject the rest. However, manual selection is also an expensive process. Thus, it is currently only practical to select records at a high level (e.g. to accept all records associated with one project and reject all records associated with another).

Longer term, it is likely that tools to help with this selection will evolve, using technologies that can assess the usefulness of potential records based on their characteristics or derived content. These tools may need to decide not just which files to keep, but also which parts of an individual file to keep. For example, imagine archiving the records created by a prominent individual. In the world of paper records, a selection decision might be taken to archive official documents but exclude, say, post-it notes. However, in the world of digital records, this distinction is potentially more difficult as both these types of records may have been sent by email and thus be stored in a single physical file.

Some current digital archives will only select records for preservation if they are created in (or first converted to) a set number of formats. This can work well if it is possible to tie the archival system to the system used to create and manage these records.

A demonstration of this principle was created by the Dutch Government's Tested project where Tessella produced an email add-in to Microsoft Outlook that allowed users to make use of their email almost as normal, but saved sent and received emails as XML documents, thereby ensuring that the format of the files

was fit for future preservation. Another example is a project to archive thesis abstracts (e.g. a Scandinavian project called DiVA uses custom-created input forms to capture the records to be archived).

However, it is reasonable to assume that data record producers will continue to use technology (and thus file formats) that match the purpose for which they are creating the record. It is thus unlikely that long-term preservation will be a major factor in this decision. Hence, a generic archival solution, although it can influence formats, will realistically be unable to impose absolute restrictions.

4.2 The structure of records

Records tend to be hierarchical in nature. The ingest process has to allow archivists to choose the most appropriate logical structure for records. For example, an accession such as the records of a committee might be logically structured as shown in figure 2.

- Records of the Committee
 - Sub-committee1
 - Agendas
 - Agenda 1
 - Agenda 2
 - ...
 - Minutes
 - Minutes 1
 - Minutes 2
 - ...
 - Sub-committee2
 - ...

Figure 2: Possible conceptual structure of a record

The records to be archived will consist of physical files, which may also have a hierarchy that will need to be preserved. For example, if a snapshot of a website is archived, it is important to preserve the folder structure so that the internal links of the site will continue to work. The physical hierarchy is more a function of current technology than conceptual content structure.

For example, the minutes of a set of meetings could be stored as:

- A series of word processor documents. In this case there is likely to be one physical file per meeting
- A dedicated database to hold all the minutes of the meetings. In this case there might be a single physical file covering the entire set of meetings
- Part of a large enterprise-wide database. In this case, the record might be a part of a series of files associated with a record at a higher level of a conceptual record hierarchy

Thus, it is important to realize that the conceptual and physical hierarchies need not be intricately linked. Therefore, it is important to allow archivists to assign a conceptual hierarchy to the records in an accession independently of the physical hierarchy of that accession. This implies that an archive should allow some files to be shared between records even if they are not associated with a higher-level record. (In this case, although the file can be retrieved as a part of any of these records, it should still only be stored once to prevent version discrepancies).

4.3 Metadata extraction

As part of the ingest process, it is necessary to ensure that appropriate metadata is captured. This metadata falls into two main broad categories: technical and descriptive.

Technical metadata allows future consumers to learn to use the records and enables archivists to take active measures to preserve them. Most of this technical metadata can be derived automatically using appropriate third-party software, e.g. to automatically determine the file format of most of the ingested files. This can then be matched against known combinations of application software, operating system and hardware that the consumer will need in order to be able to interpret these formats. This mapping can be provided by PRONOM, an on-line repository of formats created for the UK National Archives by Tessella (see section 8.1.2 below).

Descriptive metadata allows future consumers to understand the records. For paper-based records, archivists have traditionally provided manually-created, detailed descriptive metadata to accompany the records and allowed consumers

to find the records they require. It is obviously possible to do the same for digital records but the sheer quantity of these records (and the fact that it is necessary to use appropriate application software, operating systems and hardware to view them) means that this is potentially a bottleneck in the ingest process.

Ideally, this metadata would be held with the records at the point that they are created, and then updated when they are edited (e.g. through the use of a records management system). However, this is unlikely to be the case for existing digital records.

An alternative is to try to create this metadata at the point of archiving, ideally automatically. However, this is much more difficult than extracting technical metadata owing to the lack of appropriate software tools and the fact that (as described above) the relationship between a record and its physical file structure is potentially complex. Such tools would need to, for example, be capable of reliably searching through a number of files to produce a précis of their contents in a few sentences. Also there will be some contextual metadata that should be stored as part of the archival record (e.g. the background to why a certain document was produced or why it became important), which may not be contained in the actual physical files (and thus can not be extracted).

There are two potential solutions to this problem:

- Develop better automatic extraction tools. Whilst some progress is being made in this area, the field of data mining is still young and it is likely to be some time before effective tools of this nature become mature
- Rely less on such traditional finding aids for digital records and use other methods (e.g. an internet-style, content-based search engine). This option exploits one of the advantages of digital records: it is possible to search within such records, and it is a method that is becoming increasingly accepted by end users. However, even if advanced cataloguing and indexing techniques are used, it will still be preferable to offer the consumer a brief summary of the record so they can assess its potential usefulness before they are required to view, download or request the files making up that record. Also, this style of searching works well for text-based documents but is less easily applicable to images, databases etc

5. Data management

In a digital archive it is important that users are not given unauthorized access to the records. In particular no one should be able to edit the archived files. However, approved users should be able to edit the metadata about the records (e.g. to add extra information or to correct spelling mistakes). All metadata entry and editing should be audited so that it is possible to work out who changed what, and when each change occurred.

In addition:

- Many records will need to be accrued over time, so extra files must be able to be added to an existing series of records
- Migration of files should be allowed. In this case, it is advisable to always retain the original, but, if a record has undergone a series of transformations, it is not strictly necessary to retain every intermediate format. This will be discussed in more detail in section 8
- Some records will contain information that not all users will be authorized to see. In such cases, it may be possible to create a redacted version of the record consisting of files with the secure information removed. This, in essence, leads to the creation of another sibling record of the original. In general, there is no generic way of performing this redaction and thus it may be necessary to return to the original application in order to edit the files

6. Storage

One of the key aspects of an archive is to ensure that the records are stored securely and safely. Fortunately, the core of this problem is not especially challenging, as millions of organizations that retain their digital information from day to day can testify. There are, however, some unusual features needed in an archive:

- It is important to recognize the difference between the metadata (which can change) and the actual digital files (which are invariant). One way to resolve this issue is to store the metadata and files separately (e.g. in an XML-enabled database and a separate file store respectively). This means that the responsibility of keeping the links between the two

must be performed by the archive. The alternative is to store the metadata and the files together as one object, in order to ensure that they cannot become separated and to simplify backup issues. However, this can lead to the creation of unwieldy large objects, and it makes the editing of these records a potentially complicated process since it involves a large retrieval and subsequent upload, and it is then necessary to perform extra safeguards to ensure that only the metadata has been changed

- Everything that is ingested into the archive must also be backed up and stored off-site. There are a number of options for doing this but all face the same slightly unusual issue: the material to be retained is invariant. As the archive grows in size, running a full overnight backup of the system, say once a week, may not be a realistic option, and thus appropriate backup policies have to be developed that take into account the ingest rate and the relative difficulty of re-ingesting information (e.g. all information added on a given day). It is also important that the backup policy keeps the metadata and the data synchronized (if these are stored separately)
- Creating lots of copies does not by itself ensure that files are safe, hence the storage system needs to actively manage its holdings and ensure each and every file held is being appropriately cared for. This means there needs to be an automated program of checks taking place, e.g. regular checks of checksums, exercising tapes to prevent them from sticking and ensuring regular maintenance occurs before there is a problem. There are a variety of commercial systems on the market that help perform such functions
- The amount of information to be stored is often vast. One option to reduce the volume is to perform guaranteed loss-less compression. The argument against it is that this results in a loss of redundancy, so the impact of losing a data bit is high, but a key requirement of an archive should be to ensure that a single bit is never lost: and there should always be another copy of every file from which to restore
- The storage system needs to guarantee that each file has not been changed whilst it has been in its care. This can be achieved by the creation (and subsequent verification) of checksums for each file. However, this is really an application-level issue since it is best to

create a checksum quickly in the ingest process (so that the protection it offers applies as soon as a file enters the archiving system), and to verify that checksum just before dissemination. Some systems might decide to build-in digital signatures in addition to a checksum. However, the advantage of relying on just a checksum is that the technology required is simpler and openly available (checksum algorithms are freely published). The advantage of a digital signature is that it provides additional information about who verified the contents of the file, although this can also be provided by the archiving system itself and recorded in an audit trail

7. Access

Access to an archive involves searching for records and delivering them to the end user. Use of any or all of the services described in this section could incur a fee, and it would thus be the responsibility of the archive software to ensure that appropriate payment was received before allowing access to a fee-bearing service.

7.1 Finding records

There are two ways in which consumers can find records: open searching and browsing through catalogue hierarchies. The ability to perform these tasks will depend on the indexing and cataloguing that has occurred as part of the ingest process. Typically, when searching users will enter a search criteria and then be presented with a prioritized list of possible 'hits' showing a brief summary of the records that match that criteria. Users will then be able to refine the search to home in on the record they want. Alternatively, users may be able to browse through a catalogue in order to locate records.

Having chosen the record they want to see, the consumer should have the option of viewing more detailed metadata about this record (assuming this descriptive metadata has been created at ingest). As well as allowing the consumer to ensure that they have found the correct record, this information may contain important contextual information that is needed to correctly interpret the contents of the record to be retrieved. They may also be shown some technical metadata so that they can ascertain whether they are capable of physically using the record (e.g. do they need specialized application software or a long since obsolete operating system?). The issues that this latter point throws up are discussed in more detail in section 8.

When searching for records, it is quite likely that consumers would want to perform a single search to find all the available material on a given subject, regardless of whether it is stored digitally, on paper, or even in which archive it is stored (e.g. whether it is stored in a national or a regional archive). There are two ways of enabling such a search:

1. Retaining a central index containing all the relevant metadata. This option will probably give consumers the fastest response, but it means that the organization responsible for maintaining that index has to store a lot of information, ensure that it is all in compatible formats and keep it up to date
2. Distribute the search using Web services. In this case, the archive that receives the consumer's request would send out a series of sub-requests to each registered archive asking it to perform an automated search of its holdings, based on the consumer's criteria, and return a list of hits each with a numerical relevance score. It would be necessary for all the archives to agree the format of the search criteria (e.g. using an agreed XML schema) and agree a scoring scheme for the hits (although these could then be potentially weighted according to which archive they are stored in). Once all archives have replied (or a pre-specified timeout has expired) the consumer will be presented with the amalgamated and sorted list of hits. If more detailed metadata is requested on a remote holding, this could be obtained by another Web service request or by redirecting the consumer to that archive

In both cases, the user would probably need to be redirected to the archiving hosting the material for dissemination.

7.2 Disseminating records

The simplest way of disseminating records is simply to allow the users to download the records (either in their original or a migrated format). However, some downloads will be large so it may be more practical to allow users to request a posted copy (e.g. on a CD).

Both of these methods have the disadvantage of requiring consumers to have appropriate client-side application software to interpret the file formats. Thus, a third option is to create presentation-ready versions of records (e.g. converting word processing documents into HTML) and display these directly to the users.

8. Preservation planning

It might be tempting to conclude that the lesson to learn from the Domesday projects is that organizations should simply print everything that is important to them and then store the paper records or, if space is a concern, to store the information on microfiche. However, this would lose many of the potential advantages offered by digital records, such as the ability to maintain security, verify authenticity, as well as the ability to make verifiable copies, easily edit a document (if required) and search within a document. Further, for some records, such as databases or virtual reality models, it is not possible to create a printed version that captures all the relevant information. Better solutions must be found.

These solutions fall into three categories, each of which is discussed below. 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.

8.1.1 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.

8.1.2 The migration approach

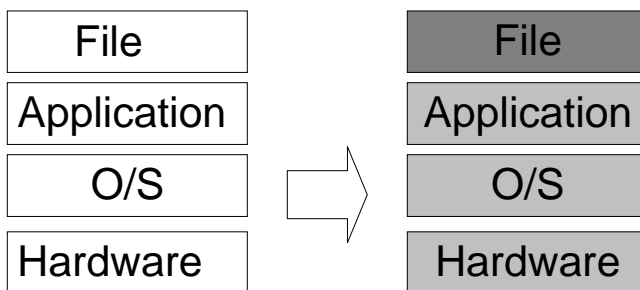


Figure 3: Migration involves accepting that natural changes occur to hardware, operating systems and application software (light grey changes) and therefore the original file is deliberately transformed (dark grey change) in order to allow a record to remain readable

In this technique, 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).

To make this approach easier, it would be best to restrict the number of formats by moving towards standardization. One example where standardization has worked is in image formats where specifications such as TIFF have been almost universally adopted by software manufacturers as they have realized that there is a larger overall market if images can be readily exchanged. In some cases, such as in colour specification, manufacturers have actively collaborated (forming the International Colour Consortium) to make standardization happen. While standardization has many attractions, the commercial companies that create the majority of application software in use today are unlikely to follow this route unless there is a competitive advantage to be gained. Also, it is worth remembering that it is not always trivial to translate records from their current formats into a standard format, as such a transformation may require archivists to make assumptions about the intentions of the original author(s).

Migration need not occur just for preservation purposes, it could also occur to allow easier presentation. For instance, if a digital record consists of Microsoft Word files, a consumer could choose to download the records to their local PC and read them using a locally-installed copy of the software. An alternative would be to create an HTML rendition of this file and display this to the consumer instead. Third-party products exist that will perform such migrations for a number of formats with a reasonable degree of integrity.

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. Behaviour. 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 (see www.nationalarchives.gov.uk/pronom/).

The ideal scenario for a large archive would be to automate the migration process. The process would then work something like this:

- An archivist updates the file format repository to state that migration of format XYZ is now required and that the approved policy is to migrate to format ABC using a specified piece of software
- The archive automatically detects the update and calculates that this will require x hours' worth of processing time
- The archive schedules this processing to occur at relatively quiet periods (e.g. over the next few nights or a weekend)
- The migration takes place automatically. Humans need only be involved to provide a quality check (although even this process could be assisted by appropriate software tools)
- An alternative to migration is to use emulation. There are variations of this technique but the most promising would seem to be hardware emulation where the original file, application software and operating system are retained but, since it is accepted that hardware will become obsolete over time, the original hardware is emulated in software on new hardware (see figure 4)

8.3.1 The emulation approach

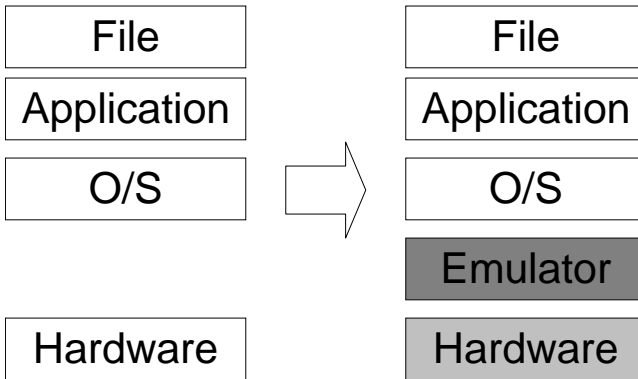


Figure 4: Hardware emulation involves accepting that natural changes occur to hardware but makes no change to the original file, application software or operating system. Enabling the software to continue to run requires the creation of an emulator (dark grey change) to emulate the original hardware on the new hardware

This technique potentially has an advantage over migration in that it should allow the look and feel and behaviour 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 behavioural 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.

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

Neither of these seem as feasible as hardware emulation (see the summary of the result of the Dutch Government Digital Preservation Testbed project for more details:

www.digitaleduurzaamheid.nl/bibliotheek/docs/white_paper_emulatie_EN.pdf).

9. Administration

Day-to-day running of a digital archive involves many tasks that are very similar to those required to keep any other large software system running. In particular, it is important to remember that all such systems require a combination of process (e.g. standard operating procedures etc.) as well as product (the actual hardware and archiving software).

However, building and maintaining a digital archive also poses some unique maintenance issues that need to be addressed during design and development. These are discussed below.

9.1 Future-proofing

The point of a digital archive is to keep digital material for a long time (potentially in perpetuity). However, one of the reasons that digital preservation is a problem is that the lifetime of most software and hardware is very small: typically just a few years before it becomes obsolete for one of a number of reasons. It is thus unrealistic to think that the software and hardware required to run the archive will last without needing significant alteration over the lifetime of the archive. To help with this, careful attention must be paid when designing an archive to provide a system that is as future-proof as possible.

9.1.1 Software issues

The first step is to use a well-established framework as a benchmark for the design of the archive. Users of archives will normally expect an interface based around Web technology (even if the archive is not released on the internet), so assuming that this is the case, an archive should be built using a standard n-tier architecture using one of the two well-established frameworks currently in operation:

- J2EE (Java 2 Enterprise Edition). This is an open standard owned by Sun. This means that application server providers using this framework (e.g. Oracle, BEA and IBM) must guarantee that their servers comply to a basic standard, although they are free to add proprietary extensions. Thus, although porting application servers is not a trivial job, it should be achievable. Most J2EE application servers are available for both Windows and Unix operating systems
- Microsoft's .NET framework. This is not an open standard, but is now well-established and is likely to continue to be supported in the foreseeable future. It can currently only be used on the Windows operating system

The following diagram (based on the Tessella design for the UK National Archives' Digital Archive) illustrates the required strict demarcation between the various application layers:

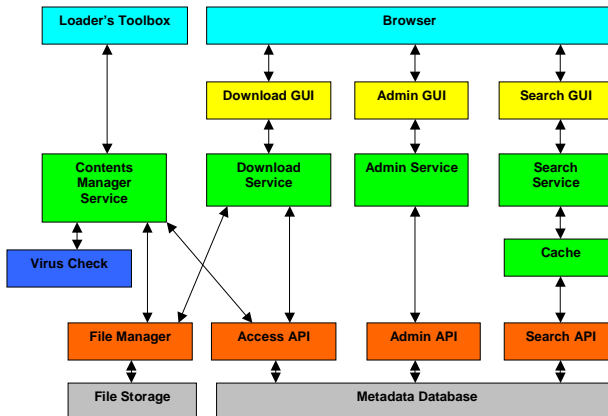


Figure 5: Tessella have designed and built a future-proof system consisting of interacting components with well-defined interfaces. This diagram shows a simplified version of this n-tier architecture

Figure 5 illustrates the second step needed to help future-proof the software in an archive: the design must be exceedingly modular with clearly defined interfaces between each component. This allows one component to be easily swapped with another without affecting the rest of the software.

A third and related feature of digital archive software is that any third-party components used must have a well-established 'sunset' policy (i.e. before any component is used it must be clear how that component could be retired and swapped with an alternative component). This can be achieved by wrapping the interface of the component in such a way that it creates a buffer between the main components of the archive and the third-party software, which insulates the former from a change in the latter. This enables a new version of the same component, or a completely different component that performs the same job, to be switched for the third-party component without affecting the rest of the application. This modular policy must also apply to the interface, with software used to control the storage of the digital files.

9.1.2 Operating system, hardware and file storage issues

Clearly an archive will need to run on real computers and thus will need to interact with real operating systems and hardware. However, there is no reason for any of the software to be tied to a particular operating system, and this should be avoided. This is one of the main reasons for choosing J2EE over .NET.

A larger issue is the hardware on which the actual files are stored. Migrating data from an old storage medium to another is a well-established technique, but the sheer quantity of material in a digital archive poses some potential problems. This means that a program of migration will need to be started well before the planned retirement date of any equipment in order to ensure that all the migrations required can occur in time.

9.1.3 Metadata storage issues

The metadata about all the records in an archive are best held in XML. This provides a format that is likely to last well into the future and will be readily understood, and is also independent of the exact nature of how it is stored, which will make migration easier. Such metadata should be captured by the user interface or imported from another source (such as record management system) and converted into an XML file compliant with the archive's XML schema. This is then likely to be appended to automatically extracted information (e.g. virus check details) as the ingest process occurs.

This metadata is probably best stored in its native XML format (e.g. by using an XML-enabled database), although some denormalization may be required to allow quick access. This metadata can then be extracted for editing purposes or converted to HTML for viewing by the user.

It is also important to have the ability to extract a full accession comprising all its XML metadata and every archived file. This functionality can be used to replicate full records between non-networked digital archive instances or to migrate the data forwards into a new system.

9.2 Procurement of tools.

As discussed in a number of sections above, there is a need for many specialist software tools to enable the effective use of digital archives. For instance, the automatic selection of records for ingest, the automatic extraction of descriptive metadata from records at ingest, and the migration from some formats to another all require high-quality software, which is not currently available, to be developed.

There are a number of ways in which this software could be procured by organizations that need to perform archival activities. Organizations could:

- Develop bespoke software themselves. This is very expensive and would result in lots of organizations repeatedly procuring similar software
- Rely on open-source software. Open-source software can solve a number of problems but it is important to realize the limitations as well. The kind of software tools needed here must be produced in a timely manner, fit into a controlled framework, be of a verifiably high quality and be supported. It is possible that open-source software can solve a few of the problems listed above but it is unrealistic to think that all the software required can be procured in this way
- Buy commercial software. The software market for these type of tools is in its infancy, but as the importance of digital preservation begins to be realized more widely, the market is likely to grow

The digital archives in existence at the time of writing have almost exclusively been built for national archives or national libraries. These institutions do not have the resources to develop their own software, nor can they afford to wait in the hope that open-source software will come along. However, one proactive role that such institutions could play is to establish benchmarks for best practice. For instance, if a piece of software is produced that performs a migration from one format to another, these benchmarks could be used to assess the ability of this software to produce a new file that faithfully reproduces various features of the original format. If such benchmark scores became recognized by the rest of the software world, as a measure of the worth of this software, it would drive manufacturers to improve the software and thus provide an ever-increasing variety of quality archival tools.

10. How can Tessella help?

Tessella has unique experience and has positioned itself at the forefront of the emerging field of Digital Preservation. This experience includes:

- The creation of the UK National Archives' Digital Archive, winner of the Pilgrim Trust and Digital Preservation Coalition 2004 Digital Preservation Award. This project involved working with the UK National Archives to specify, design, develop, test and release a system that will safeguard the UK's digital heritage for many years to come. The system is built around Java and an Oracle database. The Open Archival Information System-compatible system allows for ingest and storage of records as well as allowing access for users at the headquarters of the UK National Archives at Kew, West London. Metadata is entered (and can be subsequently edited) via a Java applet that allows users significant flexibility in entering the most appropriate record hierarchy for each accession and in associating the digital files with these records. The metadata is recorded and stored in an XML document using an e-GMS (Government Metadata Standard)-compliant metadata schema. The system is capable of handling any file type (most of which can be automatically detected via built-in third-party software). Researchers can then search and browse, view appropriate metadata, and subsequently retrieve the original records. The design has paid special attention to protecting the system against technological obsolescence so that it can evolve naturally and allow records to be kept in perpetuity. It is now a live system with an increasing number of accessions being added

- The creation of the UK National Archives' PRONOM. This system is a searchable database of file formats, related software products (for creation, migration or rendering of these formats) and their vendors. It can be used to allow archivists to plan preservation strategies. Tessella has worked with the UK National Archives to maximize the value of their investment throughout a number of project phases. PRONOM has recently been made available to users over the internet (www.nationalarchives.gov.uk/pronom/)
- The Dutch Government's Digital Preservation Testbed project evaluated possible strategies for long-term preservation of born-digital government records, leading to a set of recommendations to the Dutch government on the creation, management and long-term preservation of key electronic record types. In this multidisciplinary project team, Tessella's pivotal role was to lead the technical aspects of the work. Tessella contributed to the development of the final recommendations and designed and developed software to support the research, including a system to act as a framework for carrying out and documenting the preservation experiments and a series of preservation prototypes, e.g. for automatic metadata extraction, migration of documents between formats, and for conversion of emails and databases to XML format
- The US National Archives and Records Administration (NARA)'s Electronic Records Archives. Tessella is a partner in the Lockheed Martin-led consortium working on the project (www.archives.gov/electronic_records_archives). This is one of two consortia currently involved in the competitive design of this huge archival system
- Central Electronic Archive (CEA). Tessella helped a major pharmaceutical company create this archive to allow them to safely store a variety of data and documents needed to safeguard their intellectual property and to conform to regulatory requirements. This system deals with both archiving and disposition issues
- UKAEA JET (Joint European Torus). Tessella is a key part of the team responsible for the management of the processed data within the 17 TB store of experimental data for JET – the world's largest fusion research facility. This allows scientists to read data from decades ago, and to perform the same analysis as could have been performed at the time that the initial experimental data was produced. Tessella has been working with JET on this system for over 20 years, over which time there have been many changes of system hardware and file formats, whilst the same service to end users has been maintained

Tessella is in a unique position to work with customers to solve digital preservation problems. For further information or to discuss your potential requirements, please email info@tessella.com

Tessella - providing innovative software solutions to scientific, technical and engineering problems

Tessella specializes in the application of innovative software solutions to scientific, technical and engineering problems. Our services cover software design and development, IT consultancy, infrastructure support and project management.

Technical Supplements published by Tessella include:

- | | |
|--|--|
| <input type="checkbox"/> Active Server Pages | <input type="checkbox"/> Integrated Lab Systems |
| <input type="checkbox"/> Archiving of Electronic Information | <input type="checkbox"/> J2EE |
| <input type="checkbox"/> Automated GUI Testing | <input type="checkbox"/> Java |
| <input type="checkbox"/> Bayesian Statistics | <input type="checkbox"/> Linux |
| <input type="checkbox"/> Beowulf Clusters | <input type="checkbox"/> Microsoft .NET |
| <input type="checkbox"/> Beyond LIMS | <input type="checkbox"/> Object Oriented Programming |
| <input type="checkbox"/> C++ | <input type="checkbox"/> Open Source and Free Software |
| <input type="checkbox"/> Choosing and Using a LIMS | <input type="checkbox"/> Pocket PC |
| <input type="checkbox"/> COM | <input type="checkbox"/> Portable GUI Development |
| <input type="checkbox"/> Computational Fluid Dynamics | <input type="checkbox"/> Real Time Systems |
| <input type="checkbox"/> Computer Image Processing | <input type="checkbox"/> Regression Testing |
| <input type="checkbox"/> Decision Support Systems | <input type="checkbox"/> Security and the Internet |
| <input type="checkbox"/> Development for the Mobile Platform | <input type="checkbox"/> Simulation |
| <input type="checkbox"/> Digital Preservation: Practical Experiences | <input type="checkbox"/> Soft Computing |
| <input type="checkbox"/> e-GIF | <input type="checkbox"/> Software Development Cycle |
| <input type="checkbox"/> Electronic Data Capture | <input type="checkbox"/> Software Documentation |
| <input type="checkbox"/> Electronic Lab Notebooks | <input type="checkbox"/> Software Portability |
| <input type="checkbox"/> Evolutionary Computing | <input type="checkbox"/> Software Re-engineering |
| <input type="checkbox"/> Excel | <input type="checkbox"/> Software Specification |
| <input type="checkbox"/> Extending the Life of Software | <input type="checkbox"/> SQL |
| <input type="checkbox"/> FDA21 CFR Part 11 | <input type="checkbox"/> UNIX Inter-Process Comms |
| <input type="checkbox"/> Formulation | <input type="checkbox"/> UNIX Systems Performance |
| <input type="checkbox"/> FORTRAN90 | <input type="checkbox"/> Web Services |
| <input type="checkbox"/> Grid Computing | <input type="checkbox"/> Windows 2000 Services |
| <input type="checkbox"/> High Throughput Experimentation | <input type="checkbox"/> XML |
| <input type="checkbox"/> High Throughput Screening | <input type="checkbox"/> X Windows |
| <input type="checkbox"/> Instrumentation | |



Certificate Number FM22778



INVESTOR IN PEOPLE

Tessella Support Services plc

3 Vineyard Chambers, Abingdon, Oxordshire, OX14 3PX, England

Tel: (+44) (0) 1235 555511 Fax: (+44) (0) 1235 553301

info@tessella.com www.tessella.com