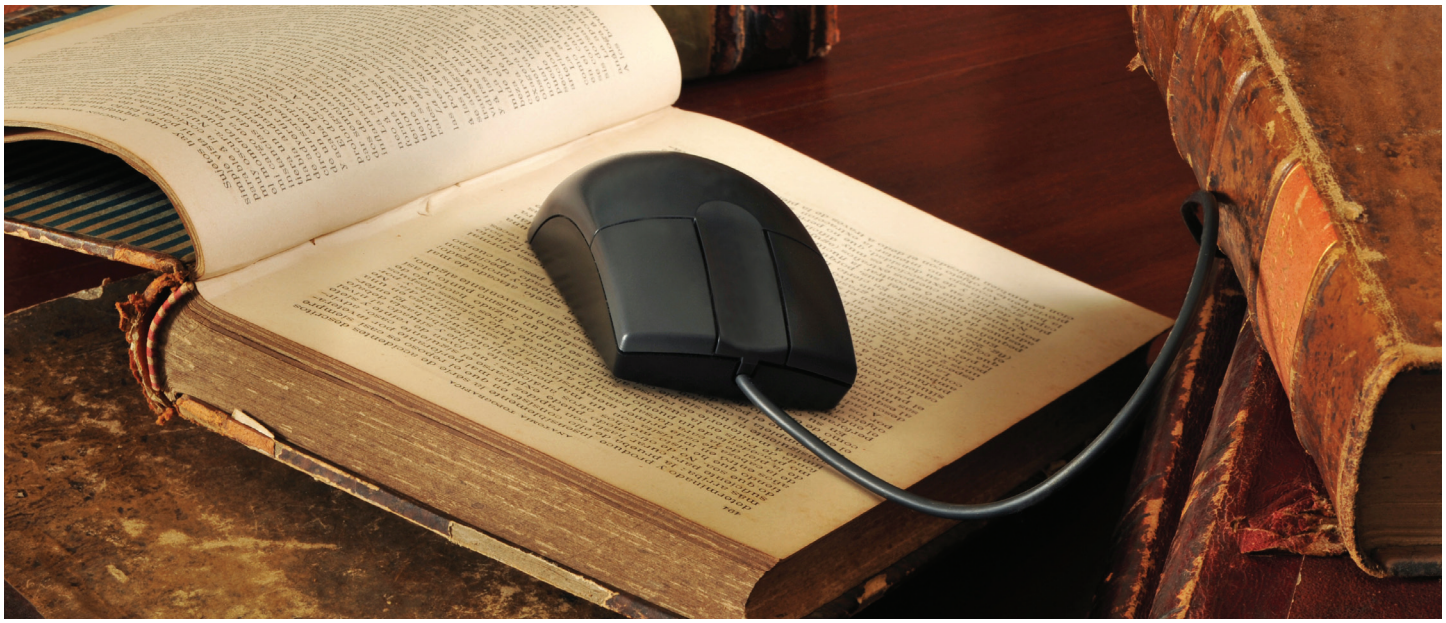
White Paper

Preservica
Active digital preservation



The Active Preservation of Digital Information

Jonathan Tilbury

Chief Executive, Preservica

Data and information management is essential within any organisation but is becoming increasingly challenging given the long and increasing time frames over which information is required to be retained. This means information contained in documents and files created many years ago, as well as those created today and in the future, is often required for time periods exceeding the supported life of the application used to create and render it. This problem shows no sign of abating, and in addition as the complexity and interconnectivity of information grows, the challenge of long-term access becomes greater.

Much pioneering work has been performed by national archives and libraries, academia and industry. In order to perform long-term digital preservation, it is necessary to (i) understand the technology of the material being stored, (ii) be able to decide whether this technology is obsolete (and if so, what to do about it) and (iii) perform verifiable actions to remove the causes of this obsolescence (for example, via format migration) or provide new approaches to delivering environments in which the original software can run (for example, via hardware emulation).

This paper covers developments in the digital preservation and archiving field based on this pioneering work. It draws on the lessons of EC-funded research projects such as Planets, KEEP and APARSEN. It also presents the lessons from national archive initiatives in the UK, Netherlands, Switzerland and beyond, and shows how these are applicable to all industries. It also shows how this experience is encapsulated in a commercially available product.

Digital preservation requires careful planning and implementation to achieve the desired output. You need to be clear why you need to retain information and what needs to be stored as a result. You need to understand how the material will be accessed and by whom, and also how long it needs to be kept and the technical challenges this might pose. The sooner these challenges are addressed, the earlier the benefits of a smoothly operating, secure, well organised and accessible archive can be realised.

Information to be retained

The motivations to retain information fall into three broad categories:

- **Regulatory.** The many local, national and international bodies that regulate industry demand that records must be kept for an appropriate length of time. The rules and the length of time retention that is required vary, but the trend is for greater retention for longer periods.
- **Legal defence.** In order to successfully defend or prosecute legal action, organisations need to present full and accurate records to the court. To be acceptable these need to provide evidence of good information management practice and be readable by those wishing to use them.
- **Knowledge re-use.** The re-use of the accumulated knowledge encapsulated in the data generated by today's activities is invaluable in advising the decision- making of tomorrow. For some organisations, for example libraries and archives, this is at the heart of their existence. In many industries this is also becoming a major driver as organisations try to recycle more value from their intellectual assets at a time when job tenure by individuals is lower and decreasing.

The information that needs to be retained of course depends on the industry. However, certain trends emerge and broad categories can be identified:

- **Operational data.** Each organisation, be it public or private, records information on its day-to-day operations. This may relate to physical activities such as plant information, laboratory data, or material movements, or intellectual activities, such as decision-making, web publication, and communications.
- **Incident data.** For any significant events, enhanced records need to be retained, including detailed instrument records, internal and public reports and management records.
- **Routine business activities.** The records relating to the running of the business including financial information, email records, quality management systems and internal information sites such as intranets and on-line forums.
- **Project data.** Any specific projects will have enhanced information that needs retaining including plans, progress and outcomes. This may be technically complex (for example, CAD drawings), or be related to other one-off activities that need to be recorded.
- **Environmental data.** All organisations have an increasing duty to record their environmental impact and for some industries, this can be of major importance. This may include assessments, instructions and sample measurements that demonstrate good practice and minimal impact.
- **Compliance data.** Additional information demonstrating compliance to the many bodies regulating each industry must be retained.
- **Human Resource information.** Staff information may need to be retained relating to attendance, health, management, remuneration and activities. These, of course, must remain confidential.

One additional complication is that this information may come from many different organisations: be they suppliers, partners, sub-contractors or customers, each of which has different practices, systems and rules. This makes implementing a simple set of rules relating to permitted formats or structures more difficult or even impossible.

The challenges of long-term access

Ensuring long-term access to this vast range of information is a huge problem that organisations are slowly waking up to. The key challenges are:

- **Media obsolescence.** Information is held on unmanaged local stores (e.g laptops, optical media, local file servers) or central stores where its value is not recognised (central file servers, tape back-ups). In each case data can be lost because the media on which it is stored becomes unreadable or is replaced.
- **Distributed and disjointed data organisation.** Another challenge is that information is stored but its location and value is not recorded making discovery and use very difficult. This challenge is compounded by information being held in diverse IT systems across several organisations.
- **File format obsolescence.** This challenge is a creation of the digital age – key information is held in files that may no longer be readable by future software. As information becomes more complex and integrated this threat is set to increase.

Media obsolescence

The issues of media failure or obsolescence have received

a lot of press coverage over the years. There are many tales of media being either unable to be read or no longer being capable of storing digital information. Some of these can be for data which cannot be re-generated, for example data from the NASA Viking Landers [1] in the 1970s or the BBC Doomsday Project [2]. This data was rescued only with large amounts of manual intervention but valuable lessons were learned and the following challenges were identified:

- **Media failure.** All forms of media are susceptible to minor or catastrophic failure. This, of course, depends on the type of media. Removable media are particularly at risk – the tapes saved many years ago may have broken or been stretched or the magnetic signal degraded.

  Optical media are also volatile. CD and DVD technology was not designed for long-term storage and the surface can be corroded in an unpredictable way. Physical damage is also highly possible. Whilst advice is available on the best approach to handle optical media [3] it is not really suitable for long-term preservation.

  Hard drives in servers are also liable to failure, and archivists constantly worry about “bit rot” and its ramifications. For all of these media types the consequences of a very small failure can be very significant. Where for paper losing one page will only result in local problems, losing just one bit in an encrypted or compressed file can result in the entire file being unreadable. At best, the results of minor bit loss are unpredictable and will require specialist intervention.
- L**ack of hardware to access media.** Removable media such as tapes, DVDs and CDs all require specific hardware to read them, and it is often surprising how quickly this becomes unavailable.

  There are tales of archivists having to go to eBay to find hardware to read media that is less than 15 years old. Part of the problem is the growing divergence within a certain technology family. For example, tapes require the specific hardware used to write them, and in optical media the variation is growing – compact disc formats now include CD-ROM, Audio CD, Video CD, CD-R (650Mb), CD-R (700Mb), CD- RW (650Mb), and CD-RW (700Mb), and for DVDs the list is longer. It is not guaranteed that the drives of the future will read all of these variations.
- **Lack of software to interpret the bits on media.** Having found some media that has not failed and read it onto a current computer system, the battle is not yet over. The bits stored on the media need to be interpreted to yield those bits generated by the application that created the file. They may have been compressed or have a particular format that needs the original algorithm to de-convolute. This may be performed separately to the access media.

Digital and disjointed data organisation

Data and documents are typically spread about in different parts of the organisation, making effective archiving very difficult. Information can be categorised as follows:

- **Raw data**. For example, this could be measured data from a laboratory, a market survey, operational notes or observations.
- **Working data**. This typically consumes the raw data and produces aggregated information such as analysis on the efficacy of a new drug or market quality.
- **Published data**. This might be a formal report for higher management or for submission to a regulatory authority.

The need for effective data organisation during the operational life of this information is really a records management issue. The problem is often that no single system is used (for example, raw data is likely to be held in files on a departmental project folder, aggregated data might be held in a database, while documents might be managed in an EDRMS system) and its production and consumption will be performed by different people in different parts of the organisation.

Depending on future usage scenarios, all of this information (and the relationships between them) might need to be maintained, leading to implications for successful archiving. Trying to build a coherent information store during end of life archiving is much harder than doing it during production.

File format obsolescence

The information described above is generated, managed, stored and distributed on computer systems. The use of paper during the active life of information is almost at an end and, increasingly, all key data is now digital. However, the range of formats this data is held in is wide and challenging. This can be summarised as follows:

- **Document based information in common formats**. Traditional records management was concerned with sets of documents that could be printed, signed, copied and distributed as paper. Electronic systems use files but the focus on document based information remains. The two main format families used today are the Microsoft Office formats: Word, Excel, Powerpoint etc, and the Adobe formats, PDF and PDF/A. Many people consider these static, that is a print of the information is an adequate representation of its content. However, this is increasingly not the case – features such as hidden text and change history make the digital copy more useful than its paper equivalent. Also, digital documents have behaviour that is important such as animations, macros and the values in a formula that are critical. Also, documents may contain embedded files that contain extra information that is not printed.
- **Specialist formats**. Much information in industry or research is held in specialist formats that are much less common. This includes CAD files containing complex 3D diagrams that cannot be printed and are used interactively on-screen. Others include the outputs from instruments and lab records, all of

which are held in digital formats that may be cross-industry but are proprietary and complex.

- **Local data formats**. Some formats are particular to specific industries, for example standard models and measurements. The use of these will be attached to software that is supported by a single organisation and may change significantly from version to version.
- **Compound data formats**. One complication that will increase as computer systems become more complex and interactive is the interrelationship between files. A CAD design for example is built up from multiple files that add layers of complexity. These files may be designed by different packages. Another example is a multimedia object which may comprise multiple video and sound streams plus text notes that are built into a single presentation. These must be used as a single unit, but managed separately. This can become more complex when a low level object, for example a common component, is used in many different places.
- **Web based information**. A web site is to some extent a good example of a series of compound data objects – each file can be made up of HTML plus images, scripts and style sheets which must be managed as a single unit. Increasingly, websites are just a representation of a database of dynamic information which is continuously changing.
- **Container formats.** A lot of data is held in containers for convenience – ZIP, TAR, GZIP etc. When considering future access it is the files within these containers that are important as well as the container itself. Some other formats such as Microsoft Office can contain embedded files making them both content files and containers. In all cases this can be multi-layered, for example a TAR file contains a ZIP file containing a MS Word file with an embedded MS Excel spreadsheet.
- **Open databases**. A great deal of information is held in databases, and many of these are “open”, that is the table names, column meanings and behaviour are understood and documented. These databases still need an application to make them useful and these applications can be highly complex, but the data itself is accessible.
- **Proprietary databases**. Much information is stored in closed databases where direct access to the data is not allowed, for example finance and HR applications or workflow systems. Access may be allowed via export systems but these can be complex to use, and the content is always changing.
- **Email**. Increasingly the backbone of corporate life, email systems can be considered a specific type of proprietary database. They contain text, attached documents and metadata relating to who sent what to who when. Often they are integrated with a corporate workflow system, e.g. MS Exchange or Lotus Notes and also refer to accreditation or certification systems that prove who sent what when.

The work of memory institutions and academia

Whilst industry has been slow to recognise these issues, the combination of archives and libraries, especially at a national level, and academia has been attacking this problem for some time. This is mainly because ensuring long-term access to knowledge is their primary reason for existence so they have given a high priority to solving this access problem before vital public knowledge is lost. Examples of their investment in this area include:

- **The Dutch Government's Digital Preservation Testbed project** [4] 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.
- **The UK National Archives PRONOM project** [5]. This is a web-based repository of information on file formats and the technical components (especially application software) needed to create or access files in such formats. This includes the freely distributable Digital Record Object Identification (DROID) tool [6], to allow files in hundreds of file formats to be appropriately identified by detecting format-specific byte sequences.
- The contribution of PRONOM to the world of digital preservation is huge. It has formed the backbone of many live systems and provides an invaluable store of factual information on data formats. Despite the National Archives' considerable ongoing investment in the content, the task is never ending and work is starting on a consortium to manage a Unified Data Format Registry based on the PRONOM data model and content. This would see major national memory institutions and academics working together to populate a single authoritative source of file format information.
- **The EU Funded Planets project** [7] is seeking to research into future ways of preserving digital information. This major project, involving 16 organisations, seeks to deliver migration technologies and validated approaches.

  The Planets approach allows organisations to characterise their content, use these outcomes to plan the best strategy and then implement this using available tools. Importantly, the outcomes are measured so that it is possible to scientifically compare strategies and tools to determine the best approach in particular circumstances.

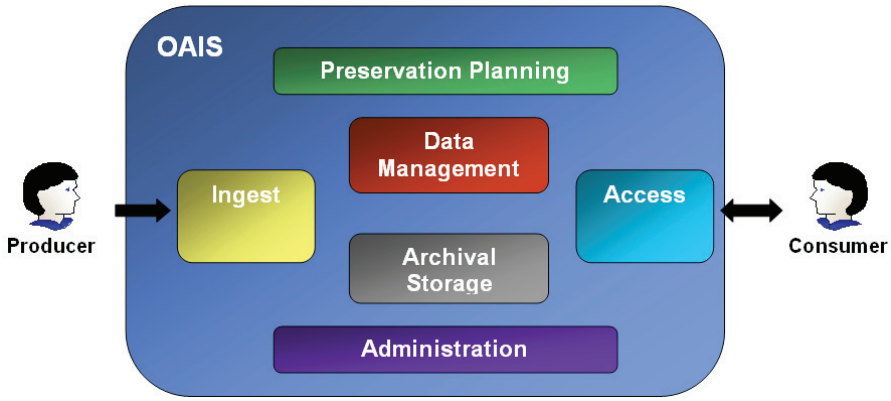
  This approach can be input as a machine-readable policy into a Technical Registry allowing automation of the entire process. Ingested content is thus characterised, compared to policy and migrated (or emulated) accordingly with the outputs measured on a case-by-case basis to ensure significant properties have not been lost in the process.
- **The EU Funded KEEP project** [8] aims to develop generic frameworks for hardware emulators. This follows on from earlier work on the Dioscuri emulator [9] for the Koninklijke Bibliotheek (Netherlands Library).

  This project is just starting but it is hoped that this will enable the automation of emulation (by rendering an obsolete object in the appropriate environment) in the same way that Planets has enabled the automation of migration tools.
- **The UK National Archives Digital Object Store**. This is one of the first general purpose digital archives which has been in production since 2003. Incorporated into the Seamless Flow programme [10], this now contains all the features of a full archive system including the active preservation of its content. This system, now available as the core of the "Safety Deposit Box" has been adopted at 9 other archives and libraries (as of March 2010).

The archives and libraries continue to invest time and intellectual effort in this area and many further developments are likely.

Open Archives Information System (OAIS) model

One of the concrete outcomes from this type of research has been the creation, initially devised by the Consultative Committee for Space Data Systems, of a standard framework for systems concerned with the management of information beyond the lifetime of the technology used to access it [11]. Now adopted as an ISO standard [12], the result is a reference model which describes the system components required for long-term preservation systems. The key elements of the model are:



- **Ingest**. These are the steps required to transfer items from their current location into the archive in a managed manner.
- **Archival Storage**. The storage of the bulk data (usually files) based on standard storage management tools.
- **Data Management**. Tools to manage the storage of the archive, including the metadata.
- **Administration**. A set of tools to administer the system and access to it.
- **Access**. Tools to search, browse and download the contents of the archive.
- **Preservation Planning**. The module that manages the information so that it can be accessed long into the future.

The use of this reference model as the basis of any archive implementation is recommended as it allows practitioners to use common language and potentially common tools to address common problems.

Implementing a basic repository

The first step to a successful long-term information management strategy is to collect the information in a central, managed location that is indexed and backed-up. This will ensure that the threats of key information stored on obsolete and volatile media are removed. The characteristics of a successful repository of this sort are:

- **Ingest**. This should make it as easy as possible to automate the gathering of information into a central location. It should also be flexible, coping with ingest from various sources using different metadata schemata, all of which will change frequently. Lack of high quality ingest tools is one of the main barriers to successful system delivery.
- **Data Management**. Tools to edit and dispose of information in a controlled manner are essential including metadata data editing, disposition to third party locations, hard and soft deletion and approval cycles.
- **Storage**. Bulk storage using commercially available file store is preferable. This may be disk only or tiered, using fast and slow disks and tape storage with a full management system. Storage should also include tools to ensure that the bytes can be retrieved exactly as saved, using checksum and, if required, electronic signing technologies.
- **Access**. Consumers of the information must have full search and browse tools with an intuitive interface allowing fast access to complex data structures. This may also be integrated with other corporate content management systems allowing fully federated search.
- **Administration**. The system must include a complex security system allowing the definition of open and closed information and role-based task allocation. Where possible this should integrate with the corporate identity management system.

Extending OAIS: Active Preservation

This type of repository is a good start but its preservation is essentially passive – it can ensure the bytes retrieved are exactly the same as those saved but cannot ensure that these bytes are usable with today's technology. A strategy to deal with this challenge must be based on "Active Preservation" – moving information to forms usable today or providing tools to allow old information to be read.

The first step in an active preservation pathway is to find out what form the information is held in. This must be done automatically – you cannot rely on the users having the time or knowhow to do this. The UK National Archives have started providing tools in this area by developing PRONOM, their database of known file formats. This integrates DROID, a tool to look at the byte patterns of the files, thus enabling format identification. The next step is to validate the file format using tools specific to the file format identified.

The next step is to extract technical information on what has been identified. This includes file-based characteristics (for example for Microsoft Word the number of pages and flags for "password protected", "contains hidden text" and "contains change history"). These characteristics might be important to determine the actions needed in preservation.

The next and more complex task is to assemble these files into the logical units of information that need to be maintained. This is an important process since files are technology-based but the real units of human-interpretable information that need to be maintained are often multi-file constructions (and indeed the number and structure of files may vary as technology moves on). For example, a GIS map could consist of many files in one technology but be aggregated into a single file in the next generation of formats. We are not interested in preserving one file per se but rather

the information. Having identified these components, their “essential characteristics” are also measured, since these are the properties that we want to ensure are maintained by future preservation actions.

At the end of this process the information is correctly identified and its characteristics are known. The next step is to identify those formats that are at risk. PRONOM provides lists of software that can read and write the formats it knows about and whether these are supported. A key risk factor is a format being held that is no longer supported. Other risk factors can include the file-based technical characteristics that were measured above, such as password protection and non-standard data structures. Using this information together it is possible to identify those files that should be targeted for preservation action.

The two key strategies currently recommended are migration and emulation. The former requires moving the data to formats currently supported, for example, moving Word 2.0 to Word 2007. Alternatively you may move it to a different format family, for example, Word 2.0 to PDF 1.4. These have their challenges, for example, Word to PDF may lose hidden text so any migration has to be validated and errors identified. This requires extracting the characteristics of the migrated file and comparing it to the original to identify changes. This can be simple (is the page count the same?) or more complex (is the image colour histogram the same?).

As described above, when migrating information it is important to move beyond the file view and migrate logical components. The best example of this occurs with web pages where migrating image files will result in broken links. It is important to follow up migrating files by migrating other files within the same logical units of information that depend on them, for example, changing the links in the HTML to point at the new image file. Also, containers must be recognised and the files within them migrated and replaced, leading to a new copy of the container file. This can lead to a cascade of migrations from one original action.

The result is a new “manifestation” of the information being managed. This terminology is important – it is not a new version as it is intended to convey the same meaning. A manifestation may combine some files that have changed and some that have not resulting in a many-to-many relationship between information objects, manifestations and files.

Emulation is a less developed approach and is intended to deliver a synthetic hardware environment on which old operating systems and programs (both preserved within the system) run to enable the original files to be used. Simulating the action of old hardware using software is complex especially where such aspects as clock speed and interaction with specific hardware are important. It can be useful for active content such as databases where the value is as much in the interaction with the data as the raw data itself. Other examples include computer games where consoles may no longer run in the future so need to be synthesised in software. As might be expected, this huge challenge requires future research to become truly useful.

The challenges discussed above extend the OAIS model to include Preservation Action as well as Preservation Planning. It is also important that as many of the actions above are automated as in very large data stores it will become too complex to migrate information individually. To that end all the identification, validation, characterisation and migration tools must be deployable at run time within a configurable system that allows actions to be run with minimum human intervention.

Learning from paper archives

The processes for paper archiving have been established over many years and are embedded into archivists' thinking. These can be summarised as:

- Bring the paper under management, and enter a proper description into an information system.
- Put it somewhere safe for storage, usually a large safe location with regulated access.
- Maintain descriptive information about the collection safely using a suitable approach.
- Enable straightforward access to the information for the consumer so they can find what they want and get access to what they are allowed to see.
- Ensure the paper is properly conserved and, if appropriate, copies are made.

This is a good description of a digital archival system as well, and indeed maps neatly onto the OAIS Reference Model. However, there are some differences that will require a change in thinking for archivists and their choice of information system should allow them to make this change:

- The process of ingesting the information tends to be much more automated, mainly due to the larger volumes handled with fewer staff. The system should make this as painless as possible.
- Safe storage is a constant issue with IT systems and without a proper policy, the risk of a catastrophic data loss remains a concern.
- IT systems tend to be rather fixed and difficult to customise. The archive should allow you to have your own processes and your own descriptive data.
- User expectations of access speeds and simplicity grown in the post-Google age means it is critical to provide fast access to appropriate information.
- Digital conservation is a complex, emerging discipline that is outside the experience of most archivists and requires specialist systems.

In moving from paper to digital it is also important to maintain a single, integrated catalogue and to choose a system that allows this to be maintained. Also, the system should allow scanned paper to be stored alongside the "born-digital" material that will increasingly dominate the collection.

Alternative digital preservation approaches

The approaches currently being followed include:

- **Backup tapes.** These are cheap and fit in with existing processes but are not searchable or easily accessible, and do not allow for format obsolescence.
- **File system.** Leaving content on an agreed location on the file system, however protected, is fraught with risk and again does not allow for format obsolescence.
- **Tiered storage.** This is cheaper than using the file system but suffers from the same risks.
- **Secure Cloud storage.** This provides online storage at a reasonable cost, especially for systems designed for delayed information return.
- **Content / Record Management Systems.** Allowing a part of the CMS to be used for archiving is a common approach. However, these management systems allow records to change, may not allow for very large volumes, do not cater for alternative content sources and have a simplistic view of format obsolescence.
- **Single content source archives.** These include specialist systems aimed at single sources of content, for example, email, finance or specific content management systems. If this is all that is needed for archiving and the retention period is short (<7 years) they can

be adequate. For longer periods or multiple content sources they start to struggle.

- **Archival information systems.** The systems currently used for managing paper may be extended to include digital media. However, they are not optimised for automated ingest, tend to have fixed workflows and metadata and do not address format obsolescence.

- **Open Source Repositories.** There are several successful open source repositories that are popular especially in academia. These have their place but often require considerable programming effort to configure, tend to be fairly fixed in what they can do, are not professionally supported and do not address format obsolescence.

- **Specialist on premise digital archives.** Specialist systems such as the Safety Deposit Box are flexible, comprehensive, supported and address format obsolescence head-on. For a dedicated archive function, they are the best solution.

- **Specialist cloud preservation as a service.** New "Software as a Service offerings" are coming on to the market which offer full Digital Preservation functionality hosted in the cloud, eliminating the need for up front capital costs and delivering a solution at a reasonable fee.

Comparing Approaches

Because of the diversity of approaches and the lack of established approach, several Digital Archives Maturity Models have been developed. One such model [14] describes the layers of a mature digital preservation system as follows:

1. **Safe Storage:** reliable bit retention on a storage media you trust

2. **Storage management:** ability to move bits between multiple storage locations to take account of durability, accessibility and cost

3. **Storage validation:** ensuring the bits are where they should be and healing from alternate copies if they are not

4. **Information management:** organising binary objects into indexed, described hierarchies of information that are accessible, secure and understandable

5. **Information processes:** the addition of integration, automation and scale to the management of the content

6. **Information preservation:** adding "Active Preservation" to make sure the files are usable by the technology available at the time they are needed.

A simple tiered storage archive would fulfil tiers 1-3 but no more. A content management archive would be between levels 4 and 5 and the specialist preservation platforms [15] [16] fulfil all 6 layers.

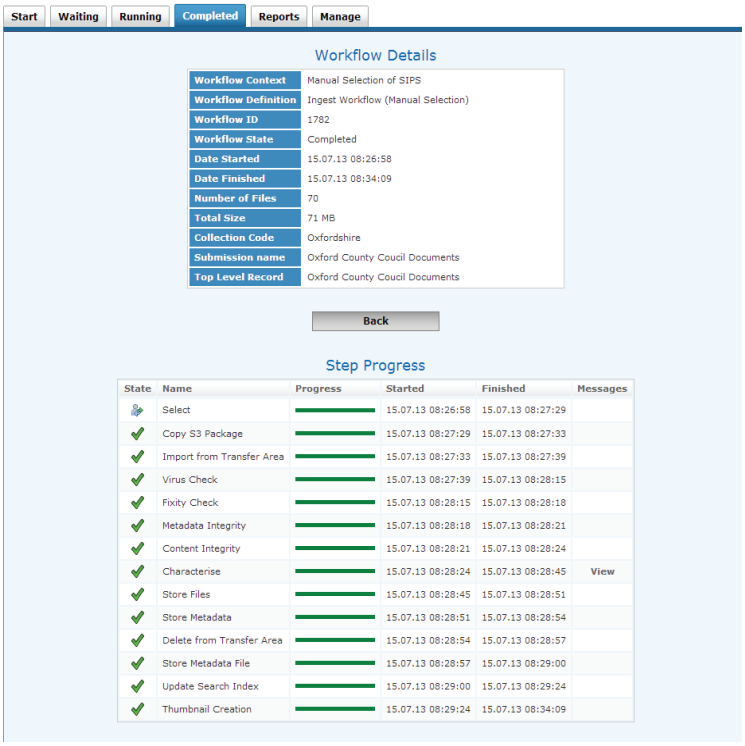
Example implementation

The original work of the UK National Archives has now been extended to create on premise and cloud-hosted Editions which are provided by Preservica, the developers of the original system. The System has the following example features:

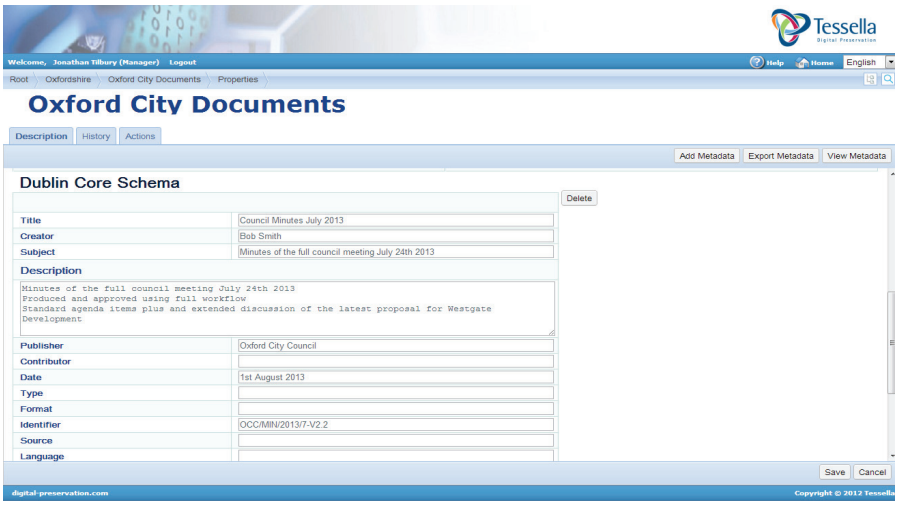
Full OAIS implementation

Functions accessible via both Java and SOAP/REST API. This allows the integration of the archive's functions into the corporate information workflows allowing an archive to be a full part of the organisation workflows.

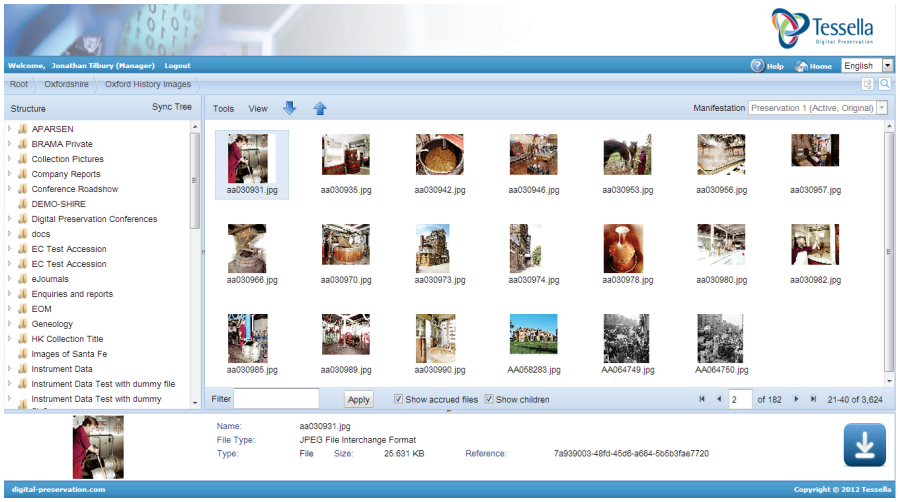
Flexible Ingest. Each organisation has a different workflow for accepting materials into its archive. This includes the steps to load the data and also the selection and assembly of materials prior to loading (the "pre-ingest" workflows). These systems implement this using a flexible Drools workflow that allows third party development of new processes and full error management.



Flexible Metadata Storage. Each organisation has its own metadata standard and sometimes several. The systems allows these descriptive schemas to be defined and used dynamically, allowing full metadata editing capability. The metadata is held in the client's chosen database system.



Access via Explorer. This allows intuitive access to search and browse via a familiar hierarchical view:



Full Active Preservation. A full implementation of the Active Preservation activities described above incorporating an enhanced version of PRONOM is included in the system. This fully automates the migration process including file and component migration.

Integration with bulk storage solutions. Preservica Enterprise Edition sits on top of proprietary bulk storage delivery and management, allowing archival processes to be linked to high quality, good value, secure storage solutions from the major global brands. Preservica Cloud Edition is implemented using infrastructure delivered by Amazon Web Services using the durable preservation storage platforms S3 and Glacier.

A strategy for whole-life information access

All sectors, whether public or private, are facing increasing demands to manage information effectively to enable controlled management and quick access. Much of the focus of this has been on the early life of the information, in email systems, content management systems and specific corporate applications. Long-term information management strategies are now recognising that the systems in place for short-term data management are not sufficient for extended time frames.

Plans for long-term information access must take into account many factors to ensure access to critical information. This starts with the initial creation and approval of the information through to its archival and long-term preservation and access. Preservation must operate at the byte level but also at the file format and information level, both of which are threats to long-term access.

Many of these challenges are not particular to a specific industry and a great deal of innovative work has been performed by national archives and libraries in partnership with academia. The OAIS model/standard and PRONOM developments have been major milestones in the field of preservation and archiving. Commercial products, such as Tessella's Safety Deposit Box, are now available incorporating this ground breaking work.

In the coming years it is expected that many more organisations across the industry spectrum will be endeavouring to meet the challenges.

Putting it into Action

The initial chapters of "Practical Digital Preservation" by Adrian Brown [17] presents a six stage plan to create and implement a Digital Preservation strategy. These are:

- **Set up a Digital Preservation mandate.** Get "in principle" agreement from stakeholders to move forward. Key first step is to understand the threats from a poor policy and the opportunities enabled by implementing a good policy.
- **Where are we now.** Understand where you are now using a Digital Asset Register which documents the different categories of information, the threats they are under, and the consequences of not having access to this information, be it a negative consequence that becomes real (e.g. regulatory penalty) or the inability to exploit a positive consequence (e.g. cannot improve business processes)
- **Where do we want to get to.** Articulate a vision of the end point in organisational terms – what will digital preservation in action look like.
- **Develop a business case.** This is one of the more difficult aspects of the process. It is greatly supported by a good Digital Asset Register, but must add a return on investment calculation and may incorporate a risk-consequence plot for each if the asset types.
- **Understand and articulate your requirements.** This formally documents the specific features you wish to implement, the stakeholders that need to be involved and the non-functional aspects of the system that must be taken into account.
- **Develop and deliver a model.** Given the growing availability of commercial and open source products this can be the more straightforward part of the process.

Once implemented, the challenge then becomes to operate and refine the approach. The growth of the collection and the enhancement of the processes will continue long into the future as the collection grows in size, diversity and usefulness.

Acknowledgments

Preservica acknowledges the innovative work conducted by the archives, libraries and universities that are members of the Planets and KEEP projects who have contributed a huge amount to the understanding of the challenges and solutions for long- term information management.

The author also acknowledges the debt of the digital archiving community to the UK National Archives in devising and populating the PRONOM database.

Bibliography

[1] NASA Viking Lander (http://nssdc.gsfc.nasa.gov/nssdc_news/sept00/viking_lander.html)

[2] BBC Doomsday Project (http://en.wikipedia.org/wiki/BBC_Domesday_Project)

[3] Handling optical media (http://www.itl.nist.gov/iad/894.05/papers/CDandDVDCareandHandlingGuide.pdf)

[4] Dutch Testbed Project (http:///www.digital-preservation.com/testbed)

[5] PRONOM (http://www.nationalarchives.gov.uk/PRONOM/Default.aspx)

[6] DROID (http://droid.sourceforge.net/)

[7] Planets Project (http://www.planets-project.eu/)

[8] KEEP Project (http://www.keep-project.eu/)

[9] Dioscuri Emulator (http://dioscuri.sourceforge.net/)

[10] Seamless Flow Programme (http://www.nationalarchives.gov.uk/electronicrecords/seamless_flow/programme.htm)

[11] ARARSEN project (http://www.alliancepermanentaccess.org/index.php/aparsen)

[12] Consultative Committee for Space Data Systems Report (http://public.ccsds.org/publications/archive/650x0b1.pdf)

[13] Space data and information transfer systems - Open archival information system - Reference model (http://www.iso.org/iso/catalogue_detail.htm?csnumber=57284)

[14] Digital Archives Maturity Model (http://www.digital-preservation.com/damm)

[15] Safety Deposit Box (http://www.digital-preservation.com/sdb)

[16] Preservica Preservation as a Service (http://www.preservca.com)

[17] Practical Digital Preservation. Adrian Brown. (Facet Publishing ISBN-13: 978-1856047555)