



DIGITAL ARCHIVING IN THE PHARMACEUTICAL INDUSTRY

Dr Robert Sharpe
TESSELLA SUPPORT SERVICES PLC

Issue V1.R1.M0
October 2006



Executive Summary

The most commonly cited reasons for retaining digital information within pharmaceutical companies is to comply with legal and regulatory obligations and to protect intellectual property. While these are important drivers, this view tends to lead to archiving being seen as a cost to the business and fails to emphasize the tremendous benefits that providing widespread access to quality information can produce. These include:

- ❑ Increasing the quality of decisions by making up-to-date information available to all who need it
- ❑ Allowing information reuse regardless of geography
- ❑ Reanalysis of old information using newer techniques or visualization techniques
- ❑ Extraction of new value from old records via data mining
- ❑ Less time wasted wondering about data sources: one search will provide it all

To achieve such benefits, it is necessary to understand the information flow within an organization as it goes through various transitions:

- ❑ Creators create data, documents etc, and initially freely use this information to enable them to perform their day-to-day tasks
- ❑ Then information passes through a phase where it more widely available, but any changes need to be controlled carefully
- ❑ Finally, only active preservation will ensure that the information remains available

At each step in this process, it is necessary to understand the value of retaining the information so that this can be assessed against the costs of doing so. The digital world has changed this perspective as the costs of storage have lowered and the potential value of the information has increased. This is because such information can now be enhanced and made widely available in efficient, summarized forms.

The last step (long-term retention of digital information) is also technically difficult owing to the fast obsolescence of the technology on which this information depends. Despite the apparent hopelessness of this situation there are practical solutions that pharmaceutical companies can employ to add real value to their businesses.

Indeed, digital preservation is maturing as a discipline. Solution frameworks have been created and developments will soon allow solutions to be assessed to see if they reach the standards required of a 'trusted digital repository'.

There are still considerable challenges to be faced; some of them unique to the pharmaceutical industry. These include the issue of long-term formats for data, the role of digital signatures, the legal admissibility of preserved documents and the wish to preserve the functionality associated with information as well as the content itself.

This paper discusses these problems, drawing on the experience of building practical archival solutions. It concludes that the technology needed to add value to the business is now sufficiently mature that the time for cost-effective action has arrived.

1. What's the problem?

1.1 Why keep data for a long time?

There are a number of reasons why digital information needs to be retained for a long time in the pharmaceutical industry. Those most commonly listed are:

- ❑ to conform with legal requirements (eg. to prevent – or help to defend against – litigation, or to comply with good governance regulations such as Sarbannes-Oxley)
- ❑ to conform with various regulatory requirements (eg. GxP and FDA 21CFR Part 11)
- ❑ to protect intellectual property (eg. to provide evidence that research has been performed and, for the current US patent system, evidence of timing)

While these are undoubtedly important, such long-term data retention needs are essentially costs to the business and do not add value themselves. However, in addition to these costs there are also the less tangible benefits of retaining digital information:

- ❑ to improve the quality of decision making by providing decision makers with access to high quality information that is easy to search and easy to access from any location around the world
- ❑ to allow information to be reused and prevent rework
- ❑ to allow information to be utilized in ways not originally envisaged when it was created (eg. by analysis using new techniques or algorithms)
- ❑ to allow the derivation of extra knowledge from a larger corpus of data than is possible from isolated results ('data mining')

If this information can be utilized to determine a new approach to addressing a disease, to help prioritize between competing projects or even to bring an unsuccessful project to an earlier conclusion, it can add considerably to the value of the organization.

1.2 Why is this difficult?

Digital information is held in files in a wide variety of formats; these can be common, such as Microsoft Word, or less common, such as a proprietary mass spectrometer output file. Each format relies on one or more pieces of application software to be able to be interpreted. This in turn relies on an operating system, which is dependent on hardware, as illustrated in Figure 1.

As a result, digital information relies on a whole stack of technical components, all of which are subject to technical obsolescence. The obsolescence of any one will render the original files to be unreadable. In addition, the files can be stored on media (eg, floppy disks, CDs, DVDs, tapes), which are also subject to rapid deterioration and/or obsolescence.

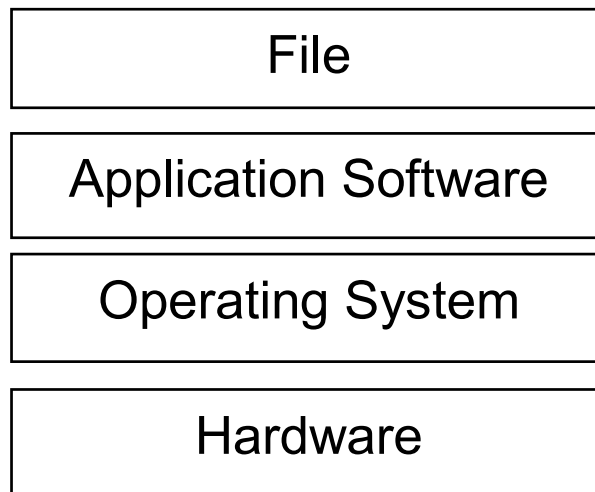


Figure 1: Digital records are dependent on a stack of technical components

The pace of change in the software industry is such that most files become difficult to access within a few years of creation. For the vast majority of information created in the world this is not an issue since the information only has short-term use: it is created, used and discarded. However, there is a significant minority that once created needs to be (or is desired to be) kept for much longer than this (e.g., electronic laboratory notebooks and the data contained therein). Indeed, some of the regulations require information to be retained for many decades, which is the equivalent to many generations of computer technology. As a result, more active preservation measures are required as part of the information flow within an organization. This is discussed in the next section.

2. Information flow

It is often not possible to decide which fraction of the information that an organization creates will have long-term value until some time after creation. This means the information retention must be an everyday part of the information flow and management structure within an organization.

This can be thought of as a three-stage process:

1. *Creation and free use*
In this stage, information is created and edited without restrictions. This could be the creation and editing of text documents, the production and quality control of data files or the everyday use of e-mail or database systems. For example, research and development staff create raw data from scientific instruments, analyze these using analytic programs or spreadsheets and prepare electronic lab notebook entries.
2. *Controlled use*
In this stage, information can still be accessed and used but any editing that occurs is controlled and audited. For example, electronic lab notebooks are searched to find a link to the raw data. Such information is non-editable but, by providing information in electronic format, it can be reanalyzed, potentially leading to the creation of new records, which need to be preserved.
3. *Active preservation*
In this stage, information is made available in new formats or in different systems than it was available originally in order to overcome the problem of technical obsolescence. For example, there is no point enabling users to retrieve old scientific data in obsolete formats: they need the same information in a format that can be analyzed and visualized in modern software.

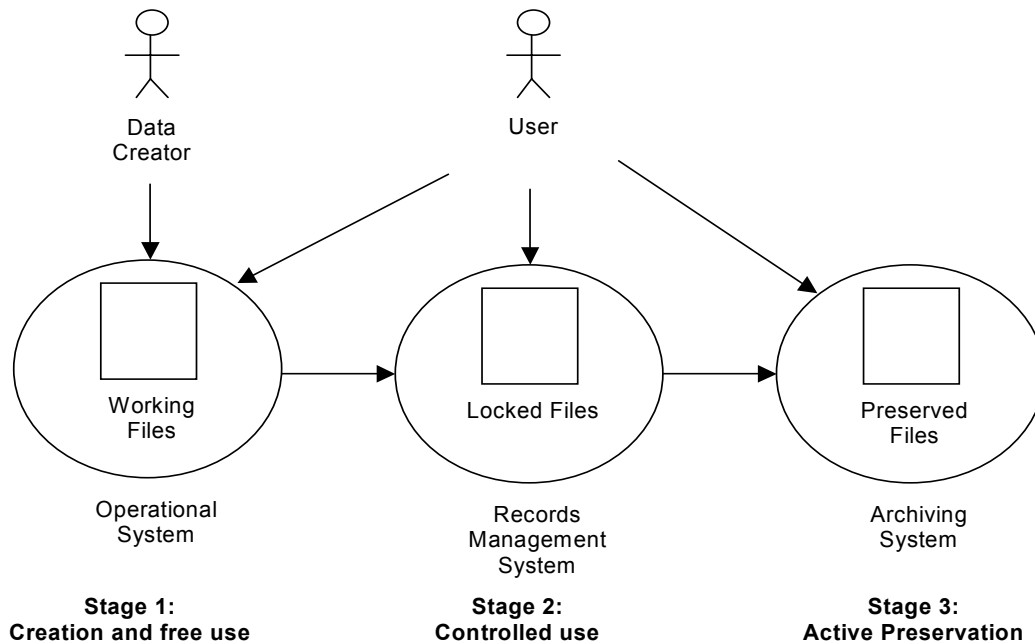


Figure 2: Information Flow within a well-managed organization

It is possible (although not essential) that each stage in this process will involve different storage systems. Regardless of the setup, it is important to have systems that smoothly transition data from active use to archive storage with minimal user effort.

2.1 Creation and free use

Information of long-term value is normally created as part of a project (eg. a new research initiative based around a treatment paradigm). However, at the time of creation, it is not clear whether such information needs to be kept longer-term, and even if it is, it is unlikely that the individual that created it (or even the project that led to it being created) will be the beneficiary of its future reuse. This means that there is a disconnect between the needs of the current user (who simply wants to create the data, consume it to draw some conclusions and then not worry about it) and the future user (who needs to be able to have the data structured so that they can find it as part of a wide-ranging search, be able to understand the context in which it was created and then is likely to want to use the data in a different manner to the way it was originally collected).

Hence it is important that data creators are able to work within a structure that does not require undue effort from them at the point of creation, but which allows information to be retained with appropriate context information so that its usefulness in the future can be assessed.

For example, it is important that, once complete, data creators store information in the correct place in a way that makes its contents clear (eg. within an application that specifies its purpose or at least by a naming convention such as storing files with an appropriate name in an appropriate folder in a project directory).

However, in many cases such structures are not well implemented and employees are free to store data files wherever they choose or in a structure which is only obvious to project team members and is quickly forgotten. For example, in many organizations (or departments within organizations) laboratory data can be left on the PCs where they were created or analyzed and are either not backed up or are stored on inappropriate long-term media (eg. CDs) in a structure that is indecipherable to anyone outside of the immediate team, and is likely to be forgotten by them within a few years.

2.2 Records Management

At some point, the initial creative burst that produced information will cease and the information will become mature. Assuming that information has been stored appropriately by its creators, the records managers should then ensure that a framework is in place to protect this information against accidental or malicious damage either by preventing changes or by allowing the editing of information via a controlled, auditable process. Again, the current reality is that such auditing is only enforced in some cases (eg. for documents that are transferred to an electronic document management system). However, the information in such documents relies on data. While it is possible to embed some of that data within a document, this is less likely to be a successful strategy for today's multi-dimensional data sets that can't be fully captured in a table of data or any other two-dimensional representation, and such a solution is unlikely to allow any further analysis that may be required at a later date. Even if systems exist for long-term data retention as well as document retention, it is unusual for these systems to be linked so that, for example, a document can make a reference to data in a way that will still allow the original data to be located decades into the future.

2.3 Archiving

In the long term, it is likely that data will be managed entirely independently of the data creators. This will probably mean moving the data to a different system while allowing the original system to get it back. The ideal would be that this occurs at set points in its lifetime (eg. a set time after its creation or last use) or is determined by a change to the original system in which it is stored (eg. it is

about to be retired or data needs to be migrated out of it for performance reasons).

These set points should allow information to be reviewed and to decide whether it is worth retaining. However, in current practice, decisions are often made simply by what is pragmatic at the point where action is needed (eg. data might be migrated to a new system only if there is a simple migration path; otherwise old data may be retained on a backup of the original system which renders access nearly impossible and on media that is liable to degradation). Indeed, in most cases it is not even known what data is missing so there is no incentive to even attempt access to what might actually be quite useful information. Hence, the next section discusses methods of deciding what to keep.

In addition, information (especially information stored in relatively rare formats such as data produced from proprietary analytical software) will be locked away in a form that is not readily accessible in many years time. The archive needs to solve this problem too, which is discussed in section 4.

3. Deciding what to keep

It has traditionally been seen as neither practical nor desirable to retain everything. However, the amount of information that can be usefully retained has increased following the development of sophisticated search techniques that allow users to find specific information from a vast data set in a reasonably short space of time (eg. the use of a tool like Google on the internet). Similarly, the development of data mining possibilities means that it is now possible to extract value from vast amounts of information where this would not previously have been possible.

Pharmaceutical companies have been leaders in the fields of records management for some time and thus, typically, already have sophisticated record retention systems especially covering signed, printed documents. Such systems allow judgements to be made about what should be retained and for how long.

They also provide processes of review and disposal at the end of such retention periods. However, modern record management needs to include digital material including not just digital documents but also forms of material with no paper equivalent such as databases, integrated data sets, Web sites, visualization and modelling outputs etc.

Setting retention periods has always been a complicated matter requiring the judgement of skilled individuals. It has traditionally been a compromise between the on-going storage costs, the cost of appraisal and the potential value of a record. The digital era, however, changes things considerably:

- ❑ Unit storage costs, in general, are much lower than the cost of retaining a paper document on a shelf – although there is a considerable up-front investment in the infrastructure and skilled personnel that are needed
- ❑ By contrast, appraisal can be much more expensive in the digital era as it is no longer possible to simply open a paper document and read it: instead, the entire technical environment discussed above (ie. hardware, operating system and application software) needs to be in place first which, for some records, eg. a database ‘dump’ from a retired system can take considerable effort
- ❑ The digital era also changes the potential value of a record by allowing greater reuse possibilities: it is easier to find information, easier and quicker to access from multiple locations rather than just the geographical site at which it happens to be stored, easier to search within records and extract partial information and easier to reanalyze or rework that information into a form that can add further value. All of these mean that information in digital form can be argued to be more valuable than the same information in paper form

In fact, the potential value of a record has always been difficult to measure. Traditionally, record retention in the pharmaceutical industry occurs only at quite a high level of abstraction: eg. by retaining only certain types of records (such as laboratory notebooks and/or high-level, aggregated result sets) and/or the project which led to the creation of the information (eg., retain the information if the project led to a product and discard if not). Indeed, recent research¹ has suggested treating archival records as intangible assets which are normally valued by knowing that a set of records is valuable without necessarily being able to calculate the value of each individual asset.

Hence, the job of the modern records manager has been considerably complicated. As a result, the sophistication of a digital archival system has to be greater than that of its paper equivalent to provide records managers and archivists with the tools needed to do their job. In the next section, we shall see how this challenge is being addressed.

4. Proposed archival solutions

4.1 Open Archival Information System

Amongst the first to realize the need to carefully look after the digital records of the future was the space science community, following some high-profile examples of the loss of exceedingly expensive data such as that gathered from Mars by NASA. This led this community to develop a reference model for the creation of 'Open Archival Information System' (OAIS), which has been adopted as an ISO standard (ISO 14721:2003)².

This work has many aspects but amongst the most important is the recognition that a long-term archive must contain at least six functional entities to allow it to operate, as shown in Figure 3:

1. *Ingest*: It must be possible to accurately move records into an archive in a form that allows all the information needed to reuse the record to be retained (including information that it may not, itself, contain such as the context of its creation)
2. *Storage*: It must be possible to store both the metadata about records and the original (and any migrated) files without losing information
3. *Access*: To be useful, appropriate users must have easy access to appropriate records. This requires user-friendly search and browse functionality
4. *Data Management*: It must be possible to upkeep information about the archive
5. *Preservation Planning*: It must be possible to decide how to maintain records within the system and implement this policy. Methods to do this are discussed more in the next section
6. *Administration*: Like all systems, an archive needs to be maintained (This is not shown in the diagram)

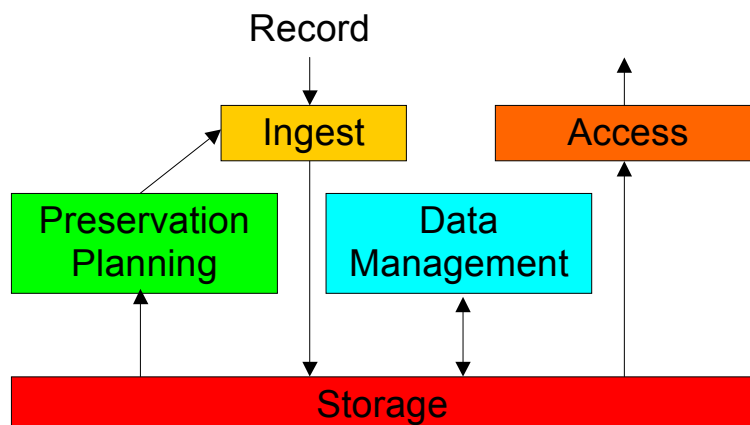


Figure 3: Schematic view of the OAIS model

4.2 Preservation Planning

There are three main approaches to preserving records:

1. The Museum Approach

Records are retained in their original form without any active preservation beyond ensuring that the bitstream is not lost. This approach cannot work in the long term owing to the issues of obsolescence discussed above. However, it is the implicit policy followed by most organizations today

2. The Emulation Approach

Records are retained in their original form and, as far as possible, their original technical environment is also retained. Clearly, it is not possible to retain the whole stack indefinitely so an emulator is created to ensure the records remain usable (eg. to allow an old operating system to run on new hardware). However, there are few examples of practical emulation in place today. Also this approach means that the original functionality of records is preserved, frozen in time. This can be an advantage but in some cases it is a disadvantage (eg. it will decrease the ability to allow the reanalysis or visualization of old scientific data using newer software or algorithms)

3. The Migration Approach

Files are deliberately changed into a new format to allow them to run on a new technical environment. This means that both the original functionality and content of records could be subject to change but, in some circumstances, this may be desirable (eg. allowing scientific data to be analyzed and/or visualized in more modern packages). As technology continues to progress, further migrations are made. It is good archival practice to also retain the original, ingested version of a record (although not necessarily any interim versions created between the original and the latest versions). This is since it allows archivists to assess authenticity by being capable, at least in principle, of returning to the original and deriving future versions from there

A good archival system should be policy-neutral, ie. allowing any of the above approaches or combinations thereof to be applied, although the reality of today's technology means that migration is the currently preferred options being followed by, for example, major national archives as is discussed below (see section 5.1).

4.3 Trusted Digital Repositories

If organizations are to commit their material into an archival system, it is imperative that the system can be trusted to retain the material. OAIS is a framework but does not tell you how to build an archive. This allows systems whether in-house or provided as a service, the potential to make good or bad decisions. How do you know whether your archive is following best practice?

To provide help with this question, the Research Libraries Group (RLG) combined with Online Computer Library Center (OCLC) in 2003 to create a guide to 'Trusted Digital Repositories: Attributes and Responsibilities'³. In 2005 RLG and the US National Archives and Records Administration (NARA) then took this further by creating 'An Audit Checklist for the Certification of Trusted Digital Repositories'⁴. Taken together these will help ensure that the efficacy of any current or proposed archival system can be assessed.

4.4 Building pharmaceutical archives

This section has so far described the frameworks in which any archival solution should operate. In section 6 we will discuss practical solutions but, first, in the next section we discuss some issues that are of particular concern to the pharmaceutical industry and thus need to be addressed in any working system.

5. Issues in the pharmaceutical industry

5.1 Formats

When performing a migration, it is desirable to be able to store information in a format that is less likely to become obsolete. As a result digital preservation practitioners around the world advocate a variety of formats for use in various circumstances. For example, PDF (or PDF/A) has been proposed for use for text documents and XML-based formats in other circumstances. Indeed, the US National Archives and Records Administration's Electronic Records Archives program⁵ proposes to create 'persistent object formats' (POFs) for a variety of data types for long-term preservation purposes.

Within the pharmaceutical sector itself there have been a number of format-related initiatives. Before the late 1980s there were few or no standards for the storage of analytical data. For example FT-IR (Fourier Transform, infrared) spectrographs were stored in a variety of proprietary formats.

Since then there have been a number of initiatives designed to bring about standardization:

- ❑ A format called JCAMP (Joint Committee on Atomic and Molecular Physical Data)⁶ was created to facilitate the exchange of infrared, NMR, mass spectrometer and ion mobility data. This has had problems in data precision and vendor-specific implementations meaning that not all JCAMP-reading software can read all JCAMP files
- ❑ The Analytical Instrument Association (AIA) created the ANalytical Data Interchange (ANDI) standards. These are based on netCDF (a common framework for storing numeric data) but suffer for a variety of drawbacks especially in allowing data changes and analysis. The standards also only cover chromatography and mass spectrometry data
- ❑ Thermo Electron SPC format. This is used in the Thermo Electron's GRAMS products primarily for spectral and chromatographic data and in many products produced by other instrument manufacturers

Lots of spectroscopic information exists in all of these 'standard' formats, as well as the proprietary formats that predate them. Indeed, the plethora of formats can be illustrated by the existence of a service to allow old spectrographs to be viewed which supports a wide variety of formats (see https://ftirsearch.com/features/converter_list.htm for a list of supported formats, many of which need to go through conversions programs before they can be read).

Since the advent of XML a new generation of self-documenting formats have emerged. An XML-based format for data was proposed by Thermo Lab Systems called GAML (Generalized Analytical Markup Language)⁷. It became a commercial standard but failed to get enough adoption for critical mass. Such formats are being replaced by AniML (Analytical Information Markup Language)⁸ but this is in its early days still.

There are also a number of other initiatives for other types of data, for example:

- ❑ ODM (Operational Data Model) from CDISC (Clinical Data Interchange Standards Consortium)⁹, which is a format in which clinical data can be stored, exchanged and archive
- ❑ DICOM¹⁰, an industry standard for the storage and transmission of medical image data
- ❑ A variety of submission formats being proposed by the FDA (Food and Drug Administration)

However, while a judicious choice of format can reduce the frequency of migration, obsolescence is inevitable as the above history of proposed formats illustrates. Similarly, XML schemas come and go from popularity and even XML itself was invented less than a decade ago with the concept of schemas considerably more recent than that. Such timescales are short compared to the proposed lifetime of many records, which may be required for many decades or even longer.

Thus, the ability of an archival system to be capable of supporting on-going, multiple migrations is more important than the particular migration policy in place at an instant of time.

5.2 What constitutes a pharmaceutical record?

In order to preserve a digital record, it is essential to store a number of things including:

- ❑ A high-level description of the content and structure of the conceptual record (ie. what is the information content and structure from a user's perspective). For example, the record could be a regulatory submission containing various documents with each document being made up of a number of constituent parts (eg. embedded data etc)
- ❑ Appropriate conceptual metadata (eg. date of submission etc)
- ❑ All the physical digital files that constitute the record
- ❑ A technical description of each file to allow preservation activities to be planned

It is also clearly necessary to be able to link the physical files and the metadata about them to the conceptual record and metadata about it. One proposed solution is the Victorian Electronic Records Strategy (VERS) approach, pioneered by the Victoria State Archive of Australia¹¹. In this approach, a VEO (VERS Encapsulated Object) is created which is a single XML document consisting of the metadata describing the record and a base-64 encoded representation of the actual files themselves. This clearly provides an intricately bound link and works very well for simple records that are made up of one or a few fairly small files.

However, pharmaceutical records can have a complex series of dependent relationships making the link between the conceptual and physical file hierarchies potentially very complex.

For example, regulatory submissions are often based on information from a variety of sources: from external CRO reports, from academic papers and from electronic lab notebook (ELN) entries. ELN entries may themselves contain embedded summaries of laboratory results. These results should be linked to the original data files of which there may be many versions: eg. raw data, quality-controlled data and summarized or aggregated data. This means that a high-level pharmaceutical record can depend on a plethora of actual physical computer files within a complex hierarchy of sub-records.

What is more, when it comes time to perform a migration, the conceptual record (ie. the user's view of the record) should remain unaltered, but the physical spread of this information between files is technology dependent and thus may be subject to change. Hence, for complex records a different approach is needed.

Tessella have built systems that allow independent (but related) management of the conceptual and physical record hierarchies. This solution:

- ❑ stores metadata in a relational database management system. This allows metadata to be added to or corrected quickly via an audited process, independently of the physical files
- ❑ delegates file storage to a hierarchical storage management system. These files do not need to be rewritten whenever a metadata change occurs. This provides a cost-effective, scalable storage solution since files can be stored in the most efficient way for their use (eg. infrequently used files can be moved onto less responsive but cheaper storage media)
- ❑ allows the system to deal with complex records such as very large Web sites that might consist of hundreds of thousands of files, where migration does not necessarily lead to a 1-to-1 correspondence between original and migrated files
- ❑ still allows fast searching for information by relying on indexes to provide search capabilities

5.3 Digital signatures and encrypted records

Digital signatures are becoming increasingly important in the pharmaceutical industry to serve two purposes:

- ❑ They identify an individual that can verify that the original version of a document (or any other type of record) was genuine
- ❑ They show that the record has not been tampered with since its original creation

Digital signatures have become commonplace and issues such as trusting signatures that originate in external bodies are being addressed through initiatives such as SAFE (Secure Access for Everyone)¹². However, today's signatures (even if standardized) rely on today's technology and it is inevitable that they will become obsolete. Hence, if we are to have confidence in the veracity of such records in the future, we need to retain both the details of the individuals and the ability to verify the signing process well into the future.

The former step can be solved (by appropriate record keeping), but the latter step requires us to rely on software that is liable to become obsolete, and thus is liable to need preservation-related attention. Fortunately, the standardization of signatures that appears to be taking place makes it likely that this problem will also be solvable.

Digital signatures do not by themselves prevent access to the files, but there is a more general problem of encrypted records (eg. anything that requires a password or another form of identification for access). There is far less standardization of encryption techniques and, while it is potentially possible to break many of these, with enough effort they do provide a considerable barrier to easy, long-term access to information.

Hence, it is necessary to give due consideration to the issues of preserving signed and encrypted records during long-term archiving. There are essentially two things that can be done:

1. In the first, records are stored in their original form (ie. signed and/or encrypted). It is then necessary to ensure that the method used for signing or encryption is well documented so that verification (subject to the identification of the verifier) or decryption is possible well into the future.

With this solution, a request for access is the trigger for both verification and /or decryption, and migration on demand to transform the original file formats into a modern form. This implies that both of these functions need to be possible for the entire lifetime of the record

2. In the second, the archive takes on the responsibility of accepting the validity of the original record. In this solution any encryption or signatures are checked on ingest and, if acceptable, removed. The archive then ensures that records are not tampered with while being held (eg. by the creation of an unsigned checksum and checking this prior to access). This approach allows the unsigned, unencrypted records to be migrated into more modern formats, as technology requires, so they are ready for immediate use upon request by a user. This also means that it is not necessary for every possible migration pathway to be active for all time so records in a given format can be migrated into new formats before the migration pathway itself becomes obsolete

Digital archival systems are in their early days so it is not yet possible to say what will emerge as best practice. However, it seems likely that some combination of the two approaches could be used, eg. following the second approach but also retaining the signed or encrypted version of the original, thereby not preventing the first method from being implemented in the future in appropriate cases (eg. to satisfy a court or regulatory body that may challenge the veracity of records held in the archive).

5.4 Legal Admissibility and migration

One of the drivers towards archiving of any type of record is that it might be needed in a court of law or by a regulatory authority. Most countries publish guidelines on the legal admissibility of digital material and in recent years, there has been an increasing acceptance by courts, at least in Western countries, that 'best evidence' laws dating back to Medieval times are out of place and thus a copy of a record is acceptable. However, it is still possible for opponents to challenge the authenticity of any evidence presented and the court reserves the right to accept or reject such evidence on a case-by-case basis. Hence, it is not possible to guarantee that archival systems will produce legally acceptable records but following certain steps will make it much more likely.

One of the issues is that courts are less likely to accept records that have been altered. However, migration (and thus alteration) is an integral part of the digital preservation process. Hence, the most important step towards legal admissibility is to ensure that the original is retained (which is good archival practice anyway, as discussed above). This will mean that, if a court is unwilling to accept a modern, migrated version of a record, it will be possible to offer the original and allow it to employ independent experts to decide whether the migrated record is indeed an accurate representation of the original.

Another important step is to ensure that any migrations that do occur are appropriately audited so that it is known who did what when and which software and tools were used.

Finally, the migration pathway and software itself needs to be assessed and validated. It is likely that any migration will lead to the loss of some information, eg. the conversion of a Word document into a PDF could lose information on previous tracked changes. Hence, it is important that any loss is characterized. The validation of the software-based processes is already well established in the pharmaceutical industry, so, in a sense, what is required is to approach migration with the same risk-based approach as that used on the processes that led to creation of information in the first place. Just as it is possible to discover that analysis and information creation software contains bugs, it will be possible to discover that migration software also contains bugs or that the migration leads to some previously unknown side-effect or data loss. Hence, the process of assessment is an on-going one with judgement-based decisions required about whether it is necessary to re-run previous migrations whenever such discoveries occur.

All of this means that the choice of migration tools and policies need careful consideration. Clearly, the creation of validated software is likely to be expensive. Fortunately, there are various schemes that allow such information and tools to be shared; in particular the PRONOM program run by the UK National Archives¹³. It will thus be possible in the future for pharmaceutical companies to follow, at least for common formats, the same procedures as national archives without having to invent their own procedures. This, hopefully, should reduce the burden of needing to research appropriate approaches and create relevant software.

5.5 Behaviour and transactional systems

Archivists discuss the properties of records and often break these down into various categories. In particular, properties are generally broken down as context, content, structure, appearance and behaviour. The first three of these properties are functions of the constituents of the actual record itself: either in the metadata held about it, in the hierarchy of the digital files that constitute the record or within the digital files themselves. In other words, they are contained within the information that is stored. However, appearance and behaviour are more complicated since these are usually functions of an interaction between the record and its original technical environment.

Behavioural attributes of records are particularly important in scientific data (where it is the analysis or visualization of data that allows information to be easily digested) and database systems. It is often thought that archiving a database is 'simply' a matter of transferring the contents of database tables into some technology-independent format (eg. an XML document) whilst recording the nature of the relationships between the fields (eg. as a script in standard SQL). However, this neglects to take into account the way in which database systems are usually designed. It is quite common for the only intended access route to a database to be through a complex n-tier system so the conceptual view of the system designed for users will go through various layers of abstraction (application server code, stored procedures, saved views and queries etc), before accessing the actual data. This means that, if the only aspects of a database system that are preserved are the table structure and contents, it is unlikely that it will be able to perform the same functions as the original system, for example the performance of complex querying and analysis.

There is currently no solution to this issue beyond treating each case on its own. NARA and the UK National Archives, amongst others, have started to address the preservation of databases and have created Access to Archival Databases (AAD)¹⁴ and National Digital Archive of Datasets (NDAD)¹⁵ respectively. In both cases these consist of database systems that have been ported on a case-by-case basis into a more generic format that can allow reasonable access to the data.

In the longer term the issue may be solved if the promise of the Semantic Web is fulfilled¹⁶. The goal of this initiative is to allow different organizations to share both data and data processing. This requires not just a universally understood expression of data but also a shared understanding of what that data means and what functions it can perform.

To this end the W3C consortium have created the Resource Description Framework (RDF), the Web Ontology language (OWL) and a standardized query language, SPARQL. It remains to be seen whether these initiatives flourish and become established but, without initiatives of this type it remains very difficult to attempt to reproduce the behaviour of records once they are removed from the system that created them.

6. Road Map

Although much of this document has highlighted issues, it is far from being all bad news. There are a number of real-life digital preservation initiatives underway. National libraries and archives, for example in the USA¹⁷, UK¹⁸, Australia¹⁹ and the Netherlands²⁰, have made great strides at building practical solutions in recent years and continue to lead the way. An advantage of these non-commercially based solutions is that they are fairly open about their progress, and thus it is possible for other types of organizations to learn. In addition, a few pharmaceutical companies are committing significant effort into the active pursuit of solutions.

Thus, the technology and the support networks needed to ensure that the value of information is maintained does now exist. As time goes by, both the current and future problems caused by poorly maintained records is only getting worse. Hence, it appears that the time is ripe for all pharmaceutical organizations to start addressing this problem.

Based on the discussions in this paper, we suggest that pharmaceutical companies should put together a road map to move from where they are currently to get to a position where information is secure into the future. The following is a suggested structure for such a road map:

- ❑ Perform an information audit that determines why information is created, why it is retained and who benefits from its retention
The aim should be complete a cost-benefit analysis so that the cost of perform archiving can be balanced against the benefits to be gained. These benefits should include both avoiding future costs (ie. regulatory compliance, IP protection etc) and the less tangible benefits to be gained from widespread access to quality information. Such analysis should enable a case to be made for appropriate changes including, potentially, the justification of creating systems (eg. to hold extremely valuable records) and the marginal benefit of utilizing such systems to

retain records that on their own would not justify such investment but are worth retaining given the unit costs of storage are lower once the system exists

- ❑ Ensure that work practices are supporting the appropriate retention of information

It is often possible to make simple changes to work practices to dramatically simplify the downstream archival process, such as ensuring that information is stored in the appropriate place or system and that metadata is assigned in a manner that provides benefit to both the current and future users. For example, end users will normally resist efforts to force them to store information with onerous quantities of metadata but a system that allows them to set a few fields (eg. project, information type) and then derives the rest of the information from stored information (eg. known properties of the project) and the context (eg. date of ingest), will not only be used but could provide these users with genuine benefits (eg. the ability to easily find their own work in a month or so)

- ❑ Ensure that electronic records management is effective

Ideally, all information that is created electronically should be retained electronically for its retention period rather than being printed, signed and filed. This requires the implementation of records management solutions. There are plenty of commercial, off-the-shelf solutions that enable this especially for documents. However, pharmaceutical companies need to come up with a solution that allows the integrated storage of both documents and data (eg. electronic lab notebook entries should be linked to raw data files). The sheer size of most pharmaceutical companies and the diversity of information that needs to be stored means that there will be a variety of record management systems required. However, to allow information links to be created (eg. from an ELN to raw data), it is important that each system should be configured to ensure that its records can be referenced by records in other systems using a unique and permanent identifier

- ❑ Create a long-term archival solution

This need not be a huge project. The archive should handle the long-term preservation of the records stored in the various records management systems in use. This means that it needs to be know of the existence of each record and the formats in the files that constitute

it. This allows it to create a proactive schedule of preservation actions (eg. migration steps) that need to be performed and to ensure that such preservation occurs using appropriate tools. In addition, the operational records management systems may delegate storage to the archive, although this is not essential if the operational system can receive and store migrated manifestations of records

7. How can Tessella help?

Tessella, the world leaders in digital archiving technology, have a range of services that can help pharmaceutical companies to exploit Digital Archiving to the full. This will enable them to benefit from the knowledge management, compliance confirmation and risk control offered by this growing and business critical field.

7.1 Consultancy services

Tessella have provided consultancy to a number of pharmaceutical companies to help them create strategies to deal with various stages of the lifecycle of information. These include practical storage options, records management, long-term digital preservation and knowledge management.

Tessella consultants can help you identify and implement the right approach for you. This will include:

- Identifying the drivers for archiving
- Categorizing the data storage needs
- Planning the data archiving workflows
- Devising a preservation strategy
- Specifying the security requirements
- Helping ensuring with compliance with legal and regulatory requirements
- Investigating how users will search and access the contents
- Planning data disposal – what happens to information once the requirement to keep it expires
- Exploring how the archive integrates with operational systems
- Selecting the appropriate software solution
- Planning and deploying the selected solution
- Advising on digitization strategies and integration with non-digital archives

7.2 Archival systems

Tessella can help pharmaceutical companies create archival systems. These can either be via integration and customization of off-the-shelf products or by customer development.

Off-the-shelf solutions very rarely provide the expected benefit right out of the box and usually require extensive customization to reflect an organization's business needs.

Tessella is highly skilled in identifying and performing this customization, given our understanding of the aims of archiving projects and a wide knowledge of software packages. We also have world leadership in implementing preservation strategies.

Many organizations have requirements that cannot be fulfilled by the software solutions currently on offer and require the software to be built for their own needs. Tessella is ideally placed to do this. We may re-use existing technologies or develop from scratch. Our years of experience in the specification, design, development, testing and roll out of major IT systems combined with our expertise in the domain will ensure the project is delivered on time, to budget and to the satisfaction of the stakeholders to provide real business benefits.

Whatever the selected archiving solution, whether your own custom software, off-the-shelf or customized third party technology, organizations only get the full benefit when it is integrated into their existing systems. This means understanding and implementing the archiving workflows, both for ingest and retrieval, allowing seamless access to the data stores.

Tessella have been involved with the creation of a number of live systems to manage long-term information. Examples include:

- ❑ Development of the Central Electronic Archive for a major pharmaceutical company 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 deposition issues
- ❑ The creation of the UK National Archives (TNA)'s Digital Archive, winner of the inaugural Pilgrim Trust and Digital Preservation Coalition Digital Preservation Award 2004. This system allows TNA's staff to ingest complex records in any format, store them securely, and provide controlled access to them. The system also allows records to be migrated, retaining the essential information content of the records whilst allowing the technology used to render that information to evolve
- ❑ Tessella is also a senior partner on the Lockheed Martin Corporation-led team that is building the US National Archives and Records Administration (NARA)'s Electronic Records Archives (ERA) system. This is a major programme that will deliver the first increment of the system in 2007 and will go on to hold a projected 320PB by 2022. The system will deal with migration of complex records to persistent object formats (POFs) in innovative ways
- ❑ Tessella is a key part of the team responsible for the management of the processed data within the 17TB store of experimental data for the Joint European Torus – the world's largest fusion research facility. This data has been generated constantly over a 20 year period and is still available live today
- ❑ Tessella has delivered a custom-built Mass Spec Data Bank, used to save experimental results at a Biotechnology company. The system must ensure that the information is saved in such a way that it can be used to support patent claims at a later date

7.3 Safety Deposit Box

Our 'Safety Deposit Box' (under development) will be a complete archiving solution ready to integrate into your operational systems or as a stand-alone solution. It will be a fully functional archive offering:

- ❑ High throughput ingest including file format recognition and virus checking
- ❑ Support for a huge range of file formats
- ❑ Advanced preservation strategies
- ❑ Ease of searching and retrieval
- ❑ Advanced security options
- ❑ Seamless integration into operational systems
- ❑ Scalability from small departmental archives to massive mass storage knowledge silos
- ❑ Safety Deposit Box will be available for direct integration into your current workflow and database systems or customization to fit the needs of your organization

7.4 Research

Although it is already possible to take a lot of practical measures to safeguard information from obsolescence, there remain significant challenges in preserving digital material for the very long term (eg. the creation of automated technology migration facilities).

Tessella's understanding of the issues surrounding digital preservation means we are able to work with organizations to research into the basic techniques required to enable the archiving to add real value. This includes proof of concept installations or individual techniques such as:

- ❑ High throughput ingest, including automatic data characterization
- ❑ Emulator technologies to allow long term access to records
- ❑ Persistent Object Formats that reduce the demands of preservation recycling
- ❑ Preservation strategies to migrate information to the latest readable form
- ❑ Integration with existing systems
- ❑ Responding to the demands of evidential proof, including digital signatures

Tessella has been involved in a number of research projects in this area, for example:

- ❑ The Dutch Government's Digital Preservation Testbed project, which 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
- ❑ Tessella is currently starting two significant projects into digital preservation technologies including emulation techniques and files transformation tools

7.5 Supporting tools

Archival systems need an array of supporting tools to allow them to provide the functionality required, such as automated characterization and migration tools. Tessella has also been involved in projects in this area:

- ❑ The development of PRONOM for the UK National Archives. 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 (see www.nationalarchives.gov.uk/pronom). The latest release also includes the freely-distributable Digital Record Object Identification (DROID), to allow files in hundreds of file formats to be appropriately identified by detecting format-specific byte sequences

For further information or to discuss your potential requirements, please email info@tessella.com

8. References

¹ Reference to work by Laurie Hunter

² <http://ssdoo.gsfc.nasa.gov/nost/isoas/>

³ www.rlg.org/en/pdfs/repositories.pdf

⁴ www.rlg.org/en/pdfs/rlgnara-repositorieschecklist.pdf

⁵ www.archives.gov/era/index.html

⁶ www.jcamp.org/

⁷ www.gaml.org/

⁸ <http://animl.sourceforge.net/>

⁹ www.xml4pharma.com/CDISC/

¹⁰ <http://medical.nema.org/>

¹¹ www.prov.vic.gov.au/vers/vers/default.htm

¹² www.safe-biopharma.org

¹³ www.nationalarchives.gov.uk/pronom/

¹⁴ <http://aad.archives.gov/aad/>

¹⁵ www.ndad.nationalarchives.gov.uk/

¹⁶ www.w3.org/2001/sw/Activity

¹⁷ See both the US National Archives and Records Administration (www.archives.gov/era/index.html) and the Library of Congress (www.digitalpreservation.gov).

¹⁸ See www.bl.uk/about/collectioncare/digpresintro.html for the British Library and, for the UK National Archives, see www.nationalarchives.gov.uk/preservation/digital.htm for current progress and www.nationalarchives.gov.uk/electronicrecords/seamless_flow/default.htm for future plans

¹⁹ Both the national library (www.nla.gov.au/initiatives/digarch.html) and archives (www.naa.gov.au/recordkeeping/preservation/digital/summary.html) are active in the field. In addition, the archives of the State of Victoria are world-renowned (www.prov.vic.gov.au/vers/vers/default.htm).

²⁰ Both the national library (<http://www.kb.nl/dnp/e-depot/e-depot-en.html>) and archive (www.digitaleduurzaamheid.nl/home.cfm) are active in the field of digital preservation

See also:

www.tessella.com/Services/Discipline/digital_preservation.htm

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.

Other 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"/> Asset Management & Monitoring | <input type="checkbox"/> Java |
| <input type="checkbox"/> Automated GUI Testing | <input type="checkbox"/> Linux |
| <input type="checkbox"/> Bayesian Statistics | <input type="checkbox"/> Microsoft .NET |
| <input type="checkbox"/> Beowulf Clusters | <input type="checkbox"/> n-tier Architecture |
| <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"/> Digital Preservation - Pharmaceutical | <input type="checkbox"/> Software Development Cycle |
| <input type="checkbox"/> e-GIF | <input type="checkbox"/> Software Documentation |
| <input type="checkbox"/> Electronic Data Capture | <input type="checkbox"/> Software Portability |
| <input type="checkbox"/> Electronic Lab Notebooks | <input type="checkbox"/> Software Re-engineering |
| <input type="checkbox"/> Evolutionary Computing | <input type="checkbox"/> Software Specification |
| <input type="checkbox"/> Excel | <input type="checkbox"/> SQL |
| <input type="checkbox"/> Extending the Life of Software | <input type="checkbox"/> UNIX Inter-Process Comms |
| <input type="checkbox"/> FDA 21 CFR Part 11 | <input type="checkbox"/> UNIX System Performance |
| <input type="checkbox"/> Formulation | <input type="checkbox"/> Web Services |
| <input type="checkbox"/> FORTRAN 90 | <input type="checkbox"/> Web Scripting |
| <input type="checkbox"/> Grid Computing | <input type="checkbox"/> Windows 2000 Services |
| <input type="checkbox"/> High Throughput Experimentation | <input type="checkbox"/> Workflow Systems |
| <input type="checkbox"/> High Throughput Screening | <input type="checkbox"/> XML |
| <input type="checkbox"/> Instrumentation | <input type="checkbox"/> X Windows |



Certificate No. FM 22778



INVESTOR IN PEOPLE

Tessella Support Services plc

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

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

info@tessella.com

www.tessella.com