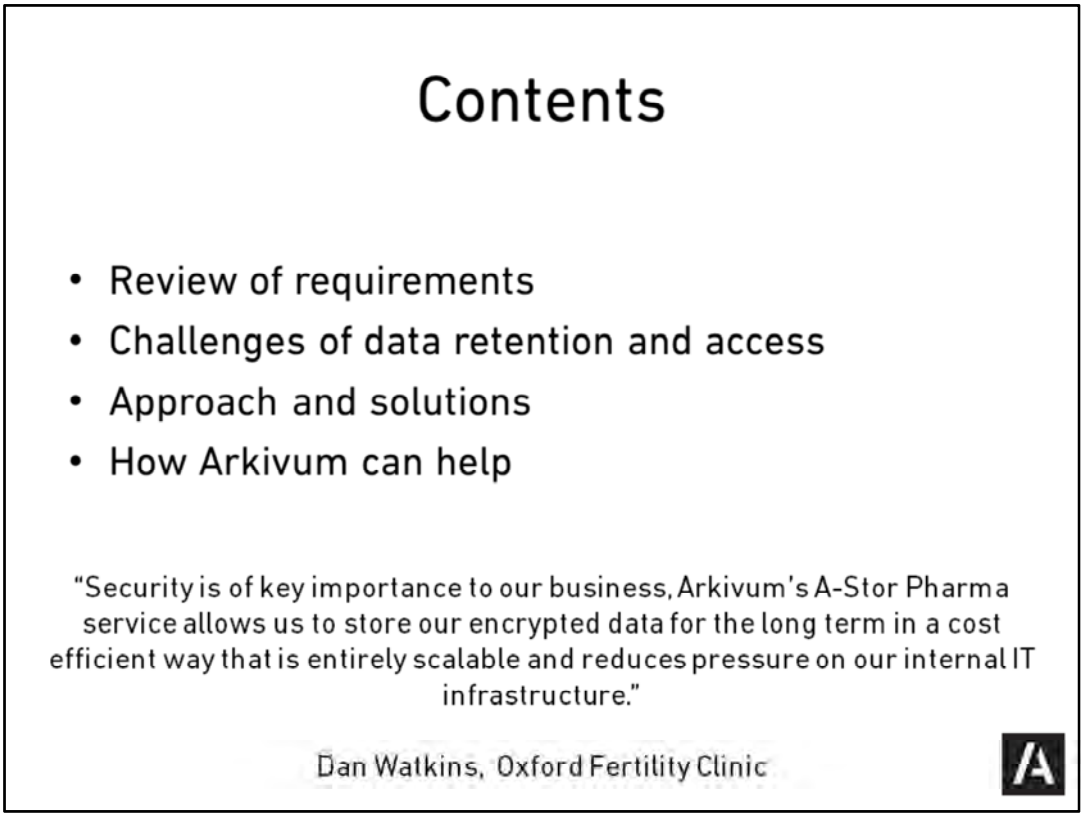
Good afternoon. Welcome to this webinar on Regulatory Compliance and Long-Term Storage of Data. It's aimed largely at lifesciences but as you'll see much of what I'm going to talk about is directly applicable to other sectors that have requirements to keep data safe, secure and usable for the long-term.

The webinar will take about 35 minutes or so, which should leave plenty of time for questions that I am very happy to take at the end.



The webinar will cover these main areas:

- The requirements driving long-term data retention and what do they mandate, especially regulatory compliance
- The challenges of long-term data retention, i.e. what's so hard about meeting the requirements
- The approaches and solutions that exist exist, especially guidelines and standards, and
- How can Arkivum provide a piece of the solution

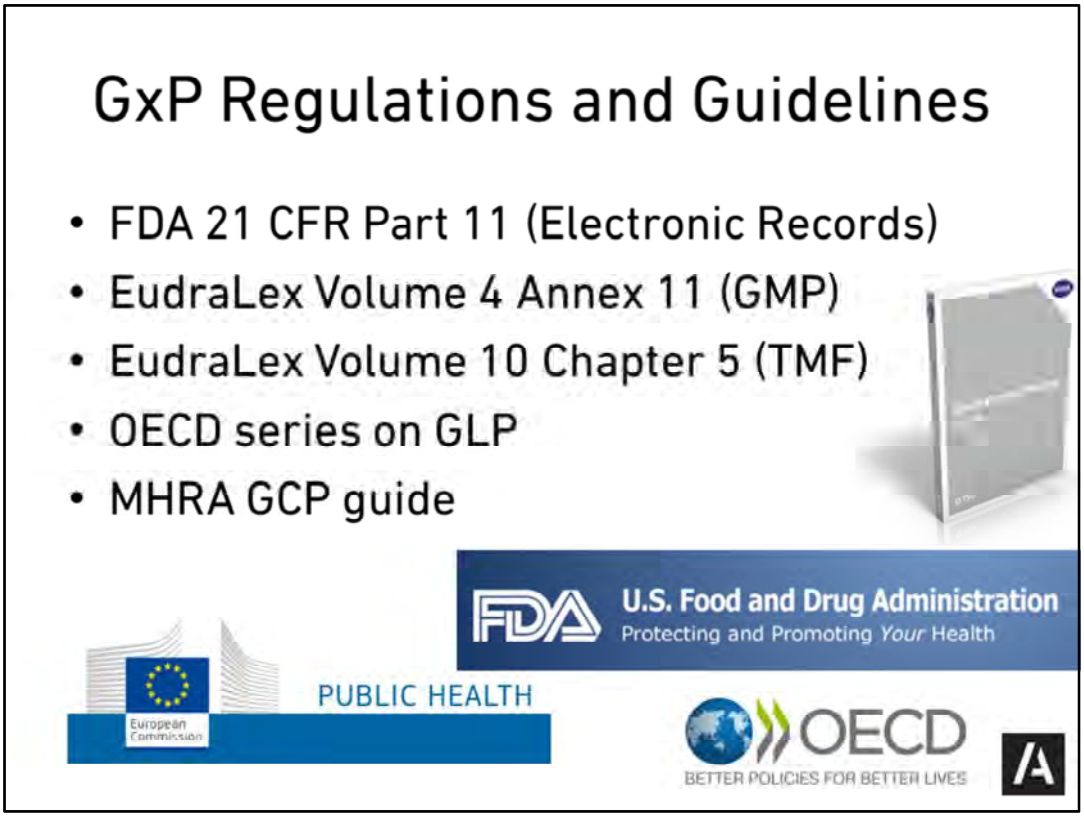
REVIEW OF REQUIREMENTS

This webinar focuses on retention of data because there are regulatory reasons to do so. But before looking at the regulatory drivers and what they mean in more detail, it's worth noting that this is just one of are several reasons to archive data. For example, in addition to regulatory compliance you might want to keep data because:

- You have value in the data, for example it supports a patent application.
- Or because you expect to reuse the data in the future, e.g. it has business value or can be monetised
- Or because it saves money. It might be a lot cheaper to keep the data than it is to create it again.

And if you have decided to keep data, then using a proper archive solution allows you to not only save money but make better use of other in-house resources. The systems used for holding data when it is first created and processed are really expensive places to store data for the long term. Moving data to a proper archive system can save money, not just by reducing primary storage needs but also in turn by needing less resource for associated maintenance and backup.

It's worth at least having these in mind as a system that supports long-term data storage for regulatory compliance may also need to fulfil some of these functions too.



So coming to some of regulatory requirements, these are useful because they structure how to think about the problem of long-term retention – as well as of course something that has to be followed.

There's 21 CFR part 11 in the US that covers electronic records, in Europe there's a range of GxP regulations e.g. Volume 4 Annex 11 and Volume 10 chapter 5. There's requirements for all stages, from GLP through to GCP – some of which are national legislation or European directives and some of which are guidelines.

What's common about these regulations and guidelines is that they talk in terms of integrity, useablity, accessibility, readability, confidentiality and responsibility – they say what needs to be achieved rather than how to achieve it.

http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?CFRPart=11
http://ec.europa.eu/health/files/eudralex/vol-4/annex11_01-2011_en.pdf
http://www.oecd.org/chemicalsafety/testing/oecdseriesonprinciplesofgoodlaboratorypracticeglpandcompliancemonitoring.htm
http://search.oecd.org/officialdocuments/displaydocumentpdf/?doclanguage=en&cote=env/jm/mono(2007)10
http://ec.europa.eu/health/documents/eudralex/vol-10/
http://ec.europa.eu/health/files/eudralex/vol-10/v10_chap5_en.pdf

Going upstream from GxP and looking at research, there is a drive towards good scientific practice that extends beyond the initial creation of research data and applies to the long-term retention and usability of that data for both the validation of the science and for the reuse of data that is created. This is why many of the funding bodies, e.g. research councils in the UK, are increasingly interested in having good data management plans in place where the quality and robustness of long-term data management matches the science that created the data. Superficially it looks like there is just a simple retention period requirement, but digging down into the details shows that the requirements are for long-term access to data that is readable and has auditable provenance and integrity, so is actually very similar to the GxP requirements.

http://www.mrc.ac.uk/Ourresearch/Ethicsresearchguidance/Researchpractice/principles_guidelines/index.htm
http://www.rcuk.ac.uk/research/Pages/DataPolicy.aspx
http://www.rcuk.ac.uk/Publications/researchers/Pages/grc.aspx
http://www.rcuk.ac.uk/documents/reviews/grc/RCUKPolicyandGuidelinesonGovernanceofGoodResearchPracticeFebruary2013.pdf
https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/206801/G8_Science_Meeting_Statement_12_June_2013.pdf
https://www.gov.uk/government/publications/open-data-charter/g8-open-data-charter-and-technical-annex
http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/projects/sape/2012-06-20-SAOE.pdf

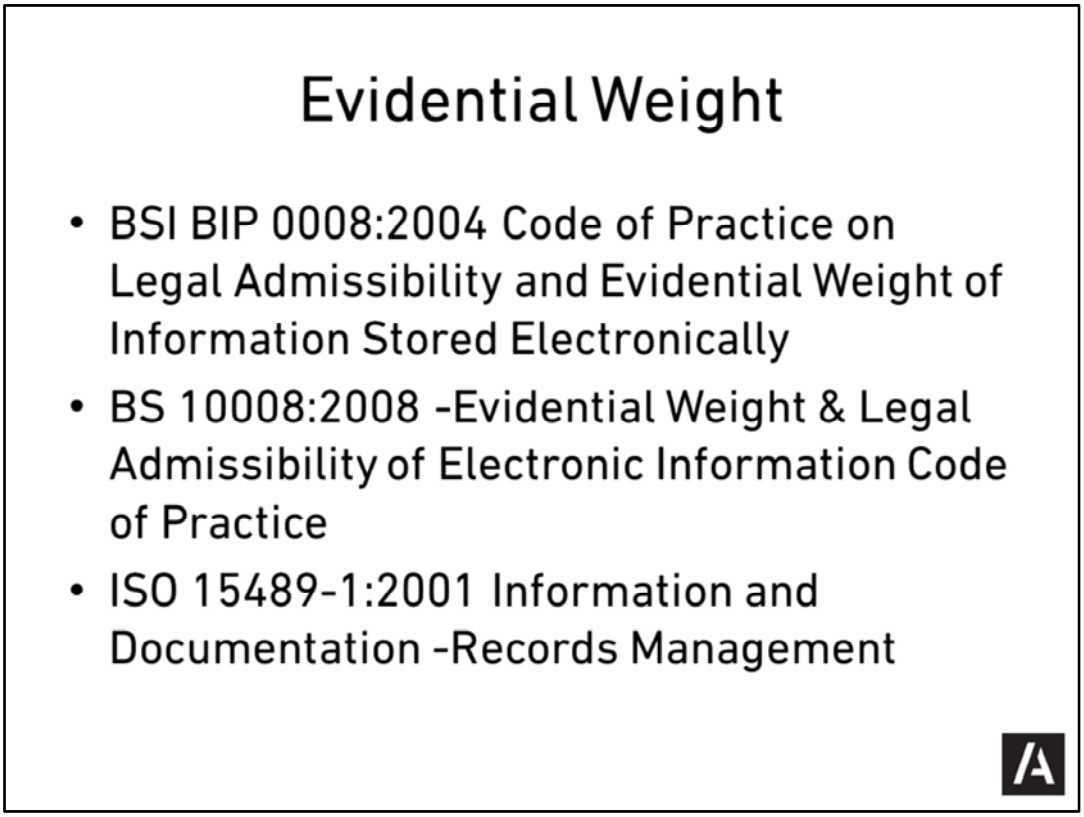
Likewise consider medical records. For example, in the UK the royal college of pathologists sets specific requirements for retention and storage of pathology records, which includes electronic records of the results of pathology investigations such as blood tests, gene screening, biopsies etc.

These records are under the broader NHS requirements for records management that covers a wide range of data, including x-rays and other types of imaging. They state that for records in digital format, then the maintenance in terms of back-up and planned migration to new platforms should be designed and scheduled to ensure continuing access to readable information

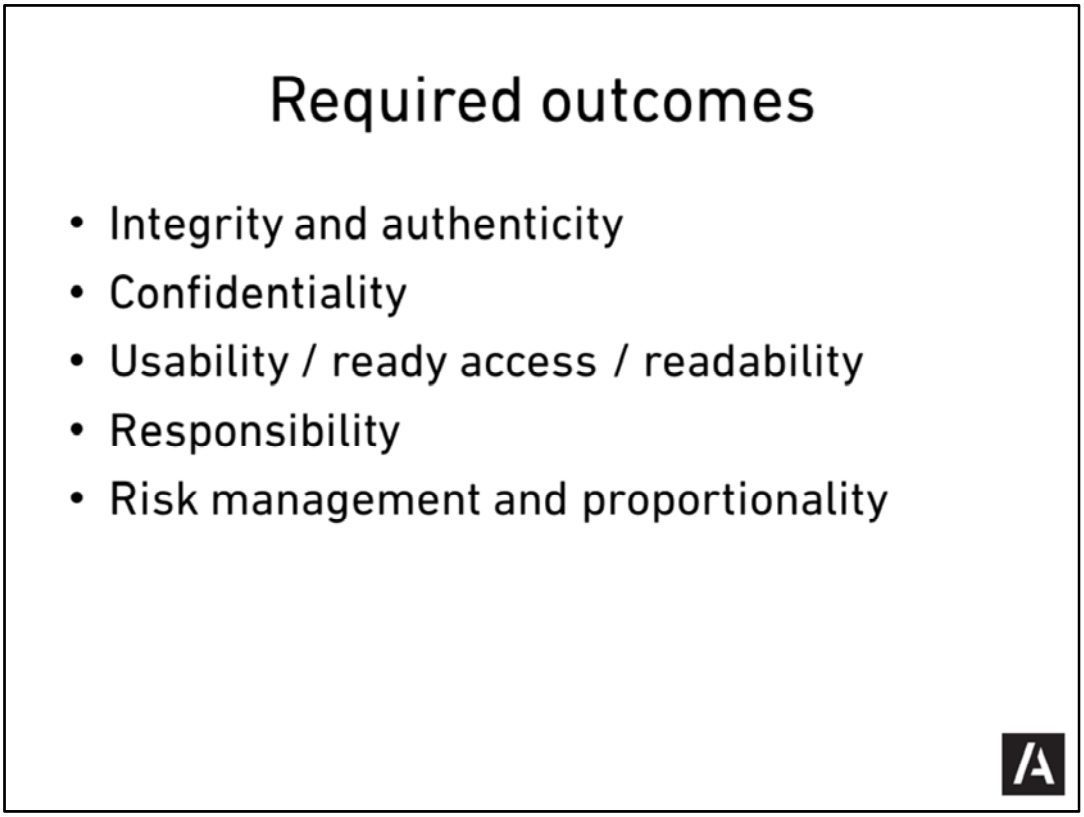
Retaining medical records is all about retention and readability but in a way that ensures integrity and auditability.

It's also worth noting that the Records Management code of practice in turn refers to audit trails and evidential weight.

https://www.gov.uk/government/publications/records-management-nhs-code-of-practice
http://www.rcpath.org/Resources/RCPath/Migrated%20Resources/Documents/G/G031RetentionandStorageAugust09.pdf



In the UK, evidential weight is covered by BS10008 and BIP0008 and these are worth a mention as they are about legal admissibility are again expressed in terms of the integrity, authenticity and availability of records. BS10008 covers policies, security, procedures, technology requirements and the auditability of electronic document management (EDM) systems. BIP0008 is a code of practice for the implementation of BS10008 and is of particular relevance where stored information may be required as evidence in legal proceedings or other disputes. It covers planning, policy, security, risk assessment, data capture and handling, monitoring, reviewing and auditing, and the maintenance and continual improvement of systems

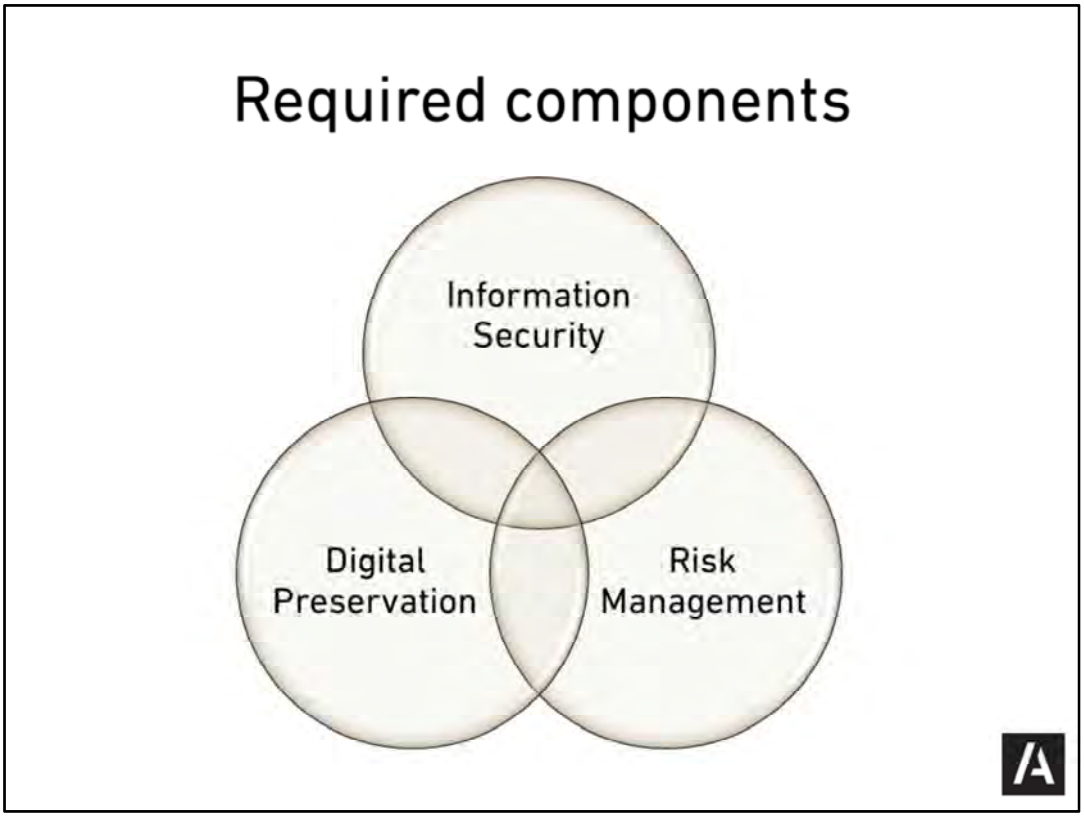


So across the board the same terminology and requirement keep coming up.

If you are keeping data for regulatory reasons, or maybe because it has value through reuse in the future, then the objectives are pretty simple.

- You need to keep content in a way that maintains its integrity, and in a way that allows you to prove this, which includes controlling access so it can't be changed.
- You need to ensure content is confidential, especially where intellectual property or personal information is included.
- You need to keep content in a way that allows you to still be able to access and understand it in the future, and
- You need to have clear responsibility assigned to individuals for the archive that holds the content

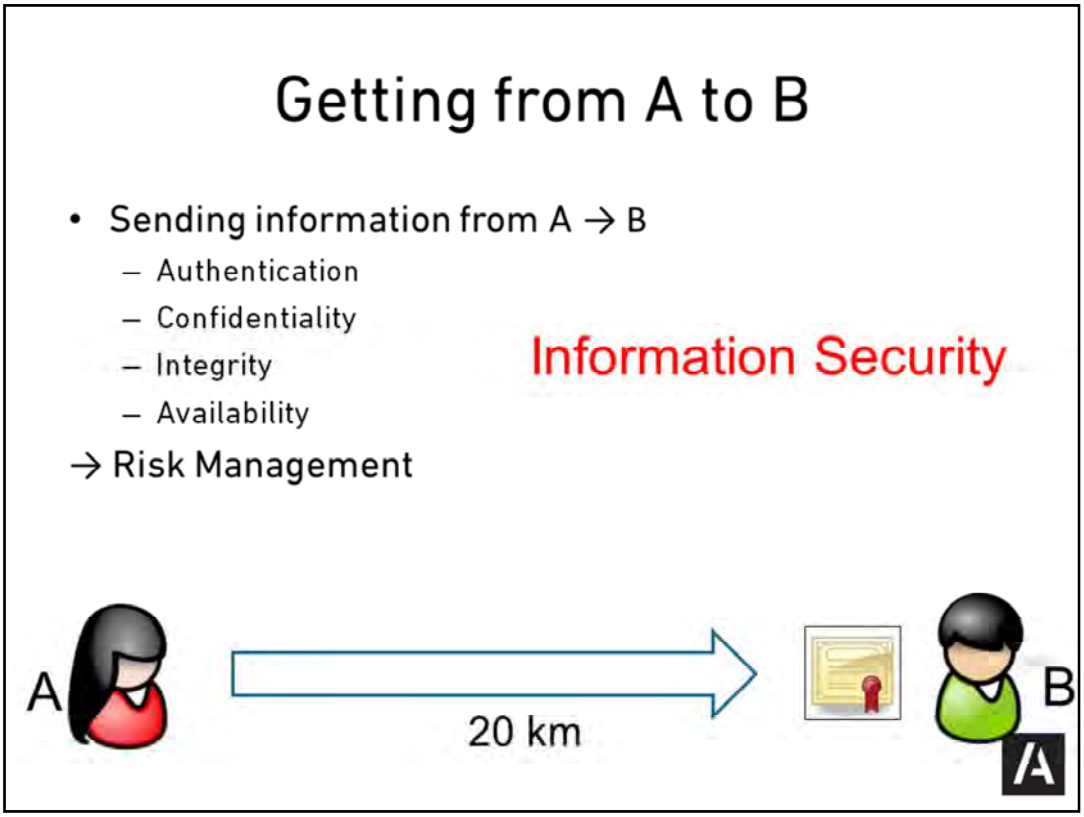
Increasingly there is a move towards a risk based approach in achieving these objectives. This means measures that are proportional to what you need to achieve in terms of safety and security – this is often implied by the regulations if not stated explicitly.



In turn this means that three main areas are required in a solution

- Information security, with its focus on integrity, availability and confidentiality
- Digital preservation , which builds on information security with its focus on accessibility and readability in the future, and
- Risk Management, which focusses on identifying managing everything that gets in the way and doing so in a proportional way

But this isn’t easy, so the next section of the webinar looks at some of the challenges.



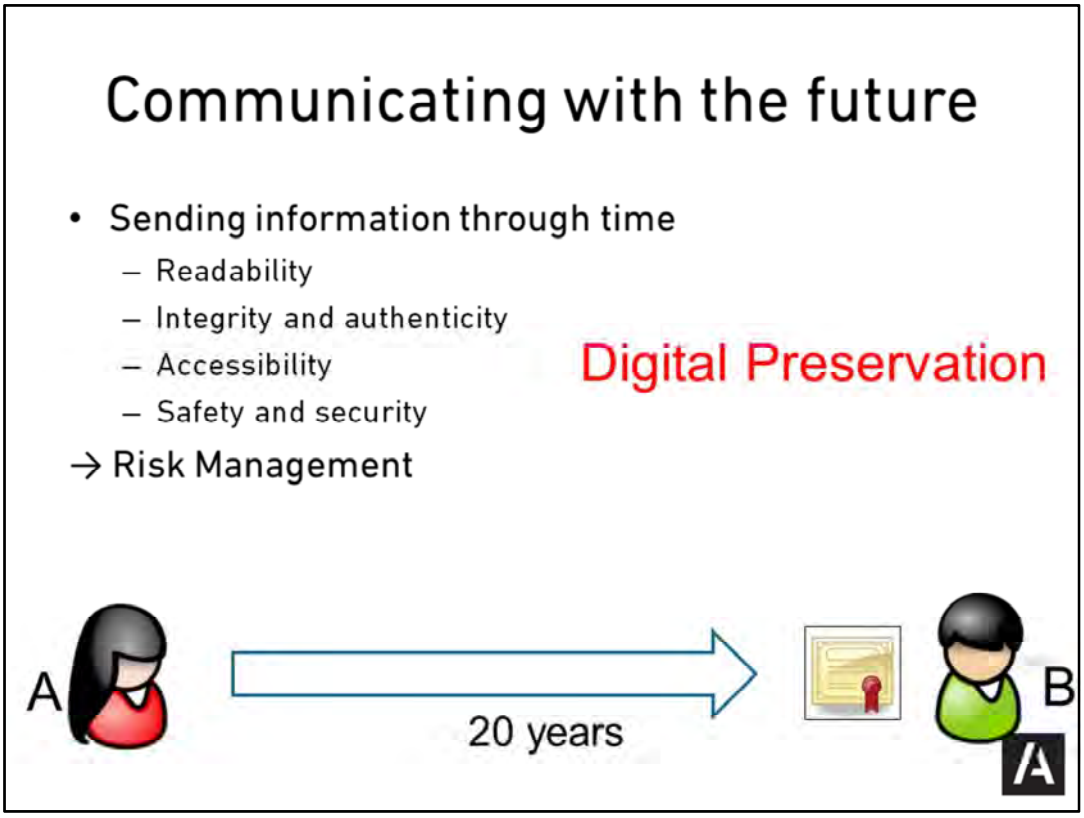
First let's think about the simple case of one person trying to send some information to another person, e.g. it could be employees of two different companies that need to exchange some valuable data.

The issues associated with this are well known and familiar to many of us.

When sending or receiving messages in a secure way:
- you need authentication, i.e. you know that the information actually came from where you think it did.
- You need to be sure that no one can get hold of the information and read it if they are not allowed to, which is about confidentiality
- You need to know that the information hasn't been tampered with or changed in some way, which is about integrity
- And you need to know that when you want to send or receive information that the system to do so will be ready when you need it, which is about availability.

The ability to send information securely from A to B is an aspect of Information Security and is standard stuff, or should be, for any IT department.

Standards like ISO27001 address information security directly and focus on having the right processes and procedures in place for managing the risks and hence this brings in the need for risk management.

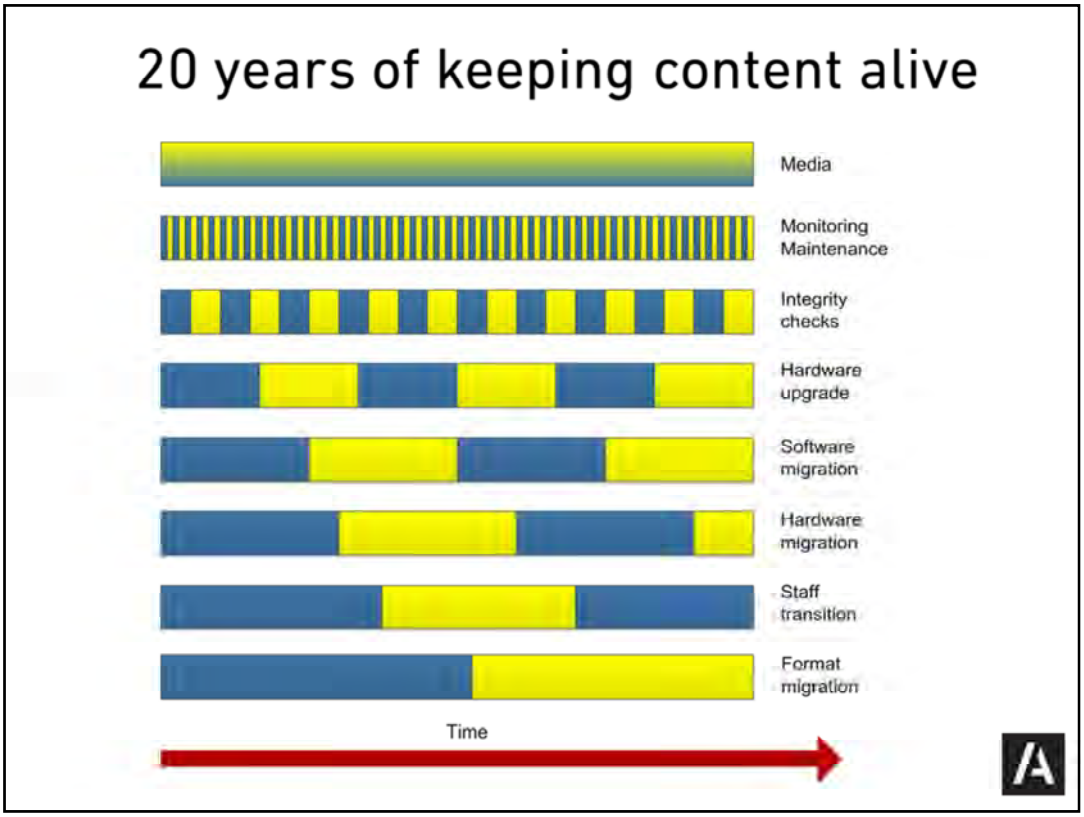


Now consider the case where instead of sending information between two people who are geographically separated you want to send the information through time. You want person B to be able to read information from person A in 20 years time.

This has all of the characteristics of information security but with the added issue of how will person B be able to understand what person A said after 20 years, i.e. is the information still readable.

This is what digital preservation is about. What makes it hard is that person B can't go back to person A and say "sorry, I wasn't able to understand that or there was a bit missing – please try again". You have to get it right first time.

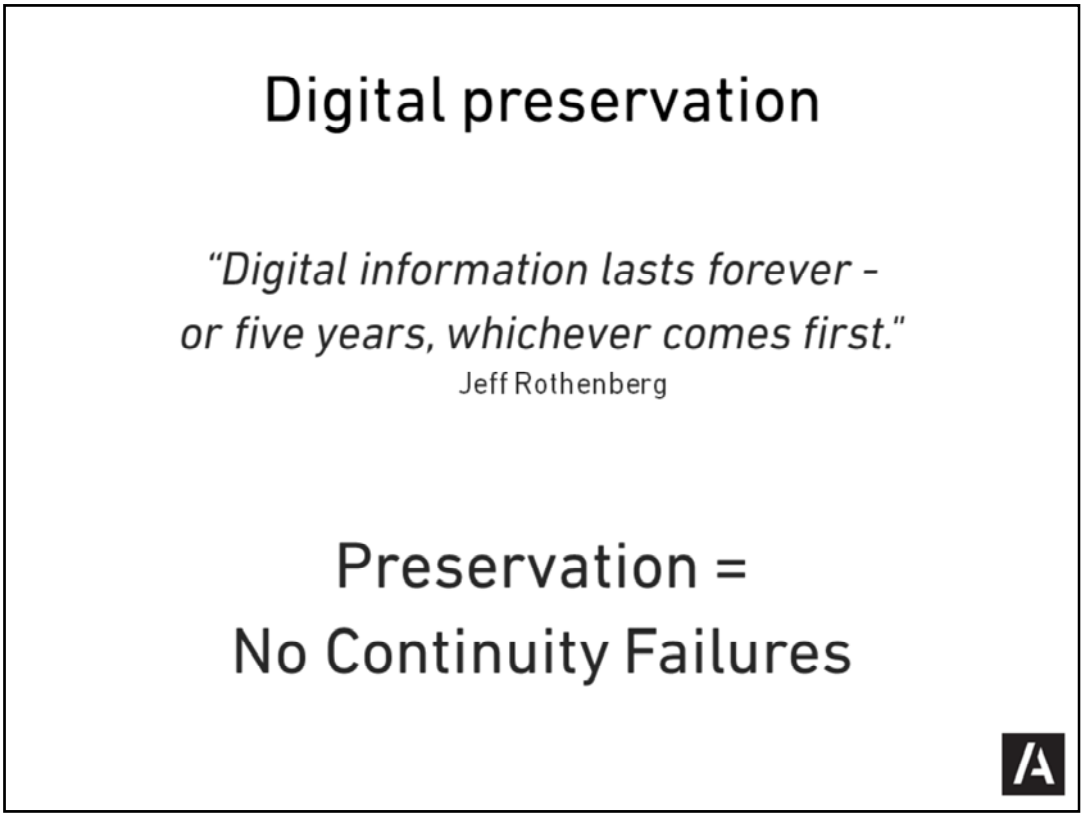
Again risk management comes into play, this time to address all the risks that happen over time.



So what are you likely to be up against? Here's a picture that shows what can happen over 20 years when you try to keep digital content safe, secure and readable.

A change from blue to yellow is when something happens that has to be managed. In a growing archive of any size, you continually need to add or replace media, e.g. tapes or discs. The archive system needs regular monitoring and maintenance, which might mean monthly checks and updates. Data integrity needs to be actively verified, for example annual retrievals and integrity tests. Then comes obsolescence of hardware and software, meaning refreshes or upgrades that will typically be every 3 – 5 years, for example servers, operating systems, application software. In addition to technical change in the archive system is managing staff transitions of those who run the system, for example support staff and administrators. Even the format of the data being held may need to change, even long-lived formats such as PDF-A will eventually be obsolete as they are replaced with something better and applications no longer provide backwards compatibility.

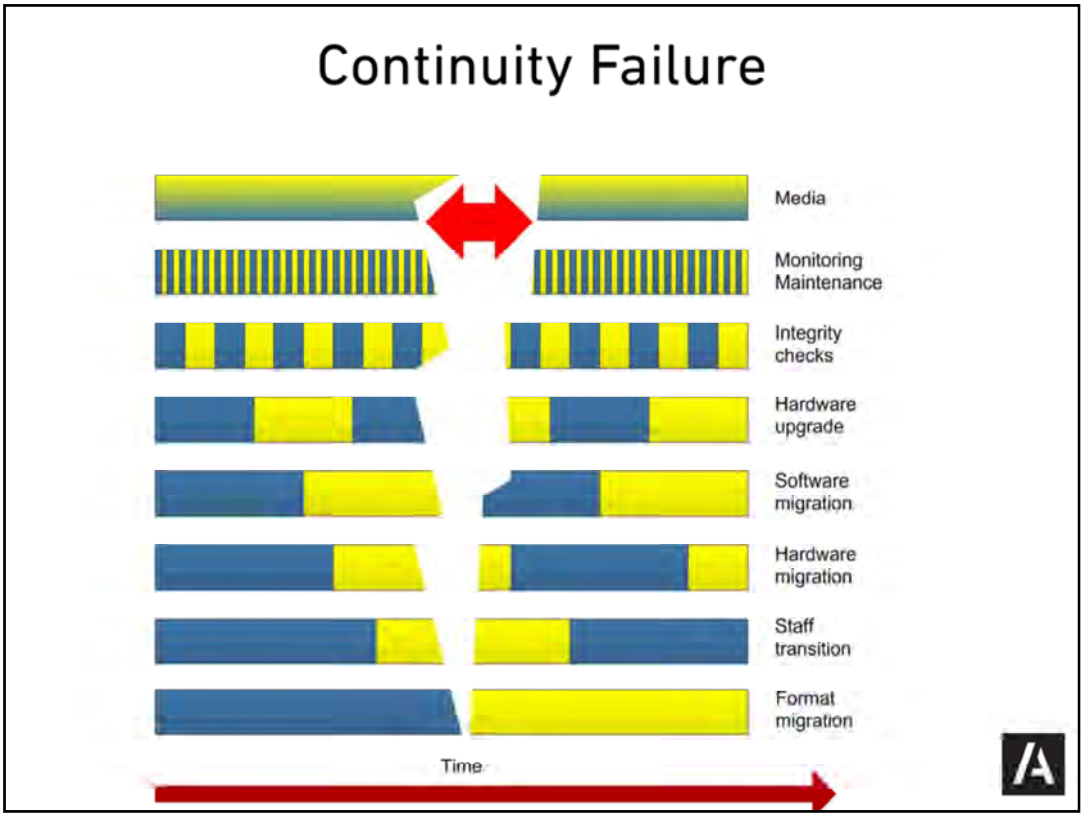
The point is that long-term archiving is an active process and there's always some form of change going on. Digital archiving is very different to paper archiving – it's not a case of 'file and forget' rather it's a case of continual interventions to keep content alive and accessible.



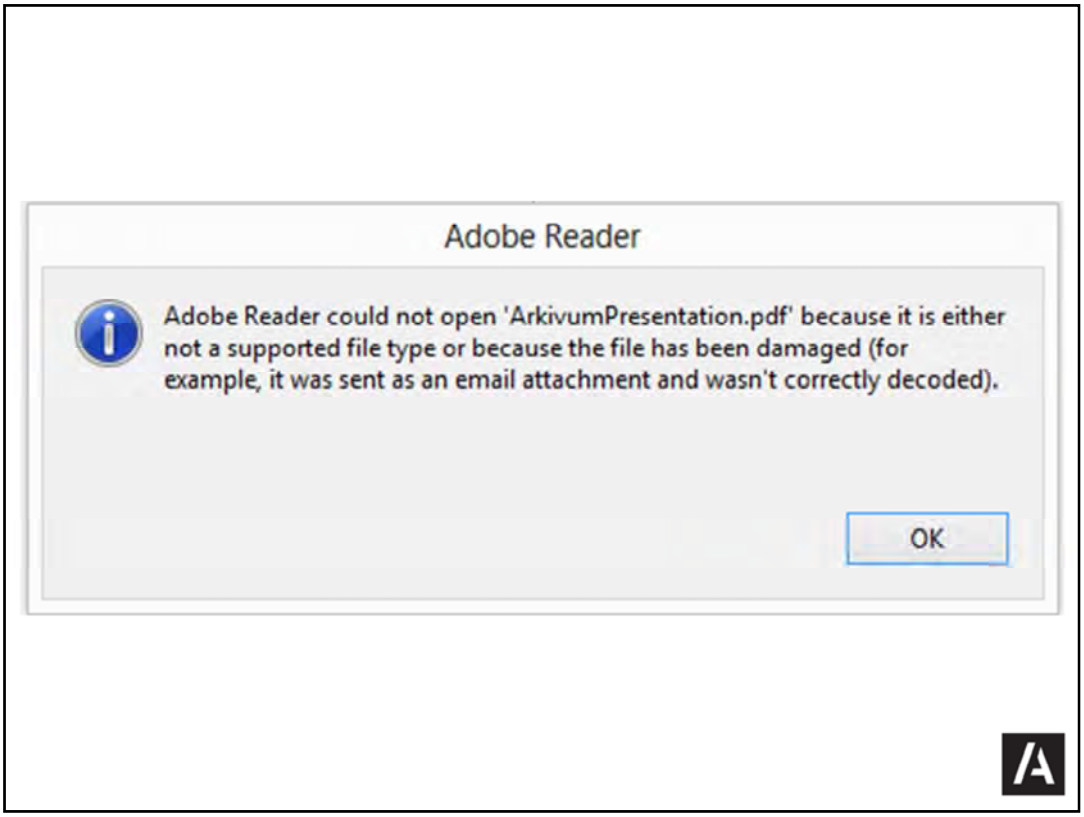
And that's why Jeff Rothenberg from the Rand Scientific Corporation, coined the phrase that digital information lasts forever or 5 years, whichever comes first, because unless you take active steps to keep digital information alive then it'll be gone.

What preservation means is not having any continuity failures – all those blue and yellow transitions have to work – otherwise the chain is broken, continuity fails, and data is at best at risk and at worst permanently lost.

[1] http://www.dpconline.org/advice/preservationhandbook/introduction/definetions-and-concepts?q=definitions
[2] http://www.clir.org/pubs/archives/ensuring.pdf

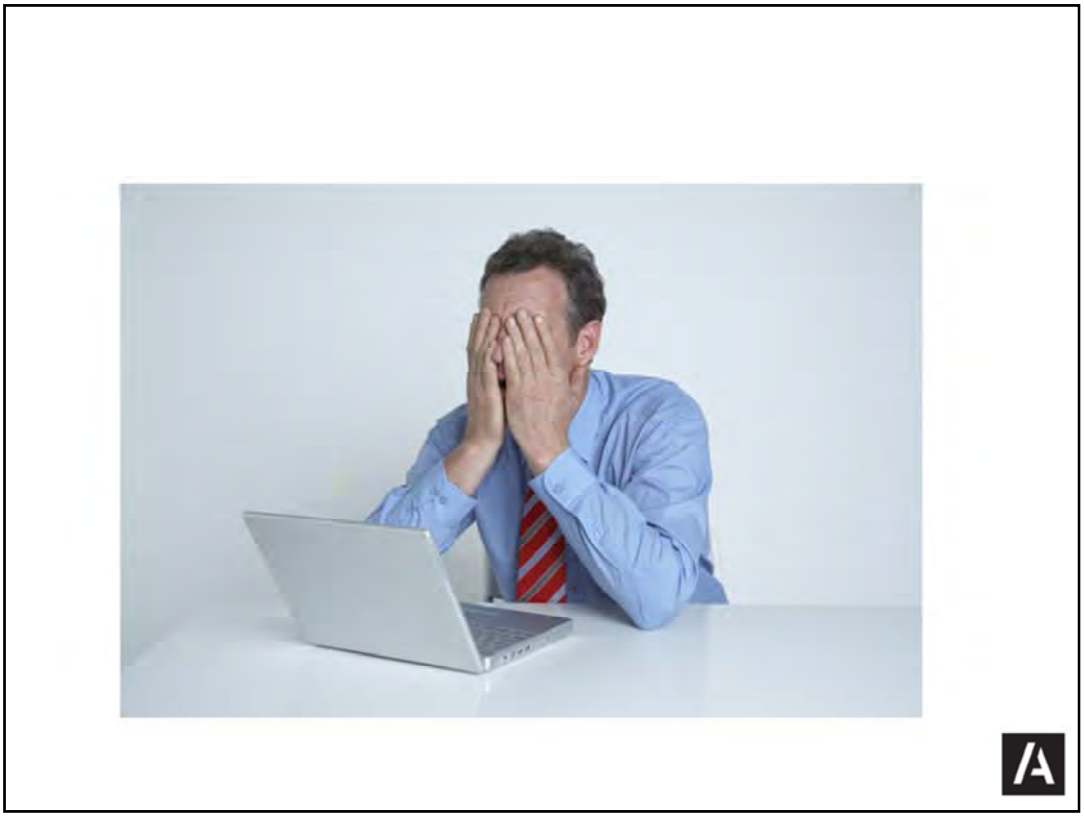


And once continuity is lost at some point then the cracks can spread and other failures creep in.



And this is what happens when there's a problem.

It's typically 'all or nothing' when it comes to reading a damaged file. You try to open a file (if its still there) and the application you're using typically throws it's hands in the air and says 'sorry, it's gone – nothing I can do'.



Which leads very quickly to this.

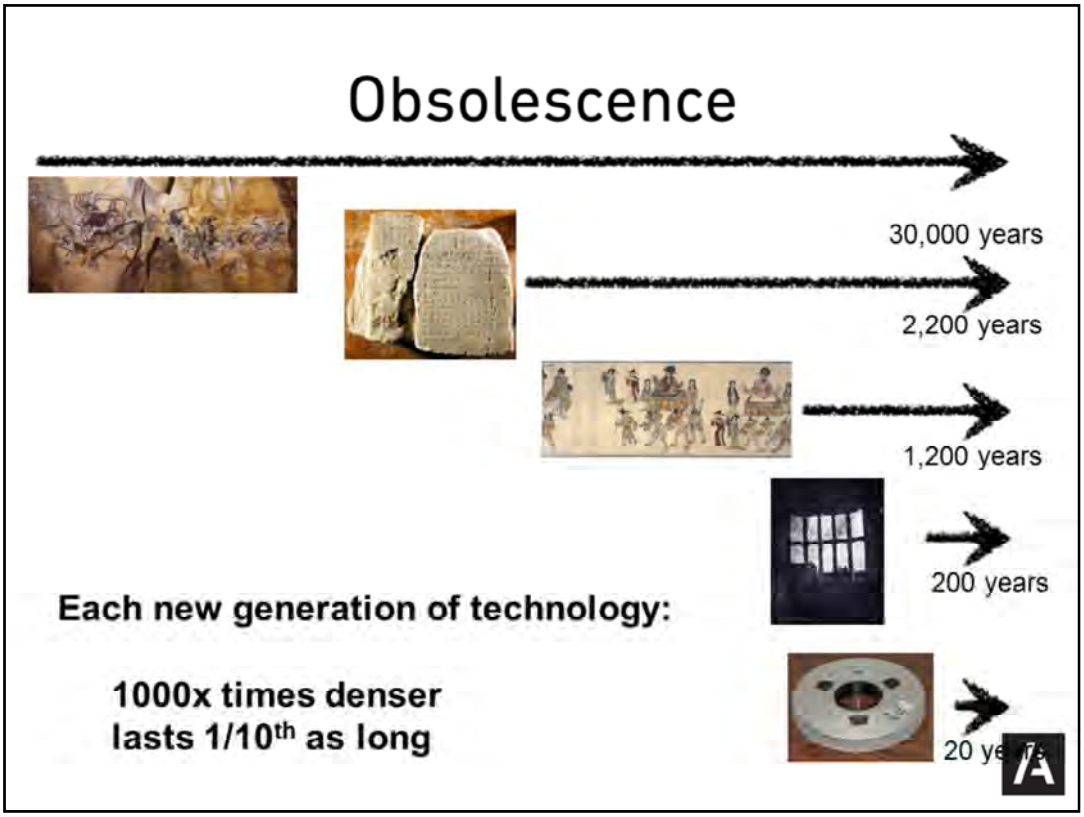
Which can then lead to fines, loss of business, and indeed loss of staff!

Continuity

- Continuity costs money
- Continuity takes time
- Continuity introduces risk
- Continuity requires validation
- Continuity needs planning
- Continuity needs management

Active steps are needed to stop this happening, which means continual intervention and continual change. It's about those blue to yellow transitions.

Change costs money
Change takes time
Change introduces risk
Change requires validation, especially in a regulated environment
Change needs planning and management



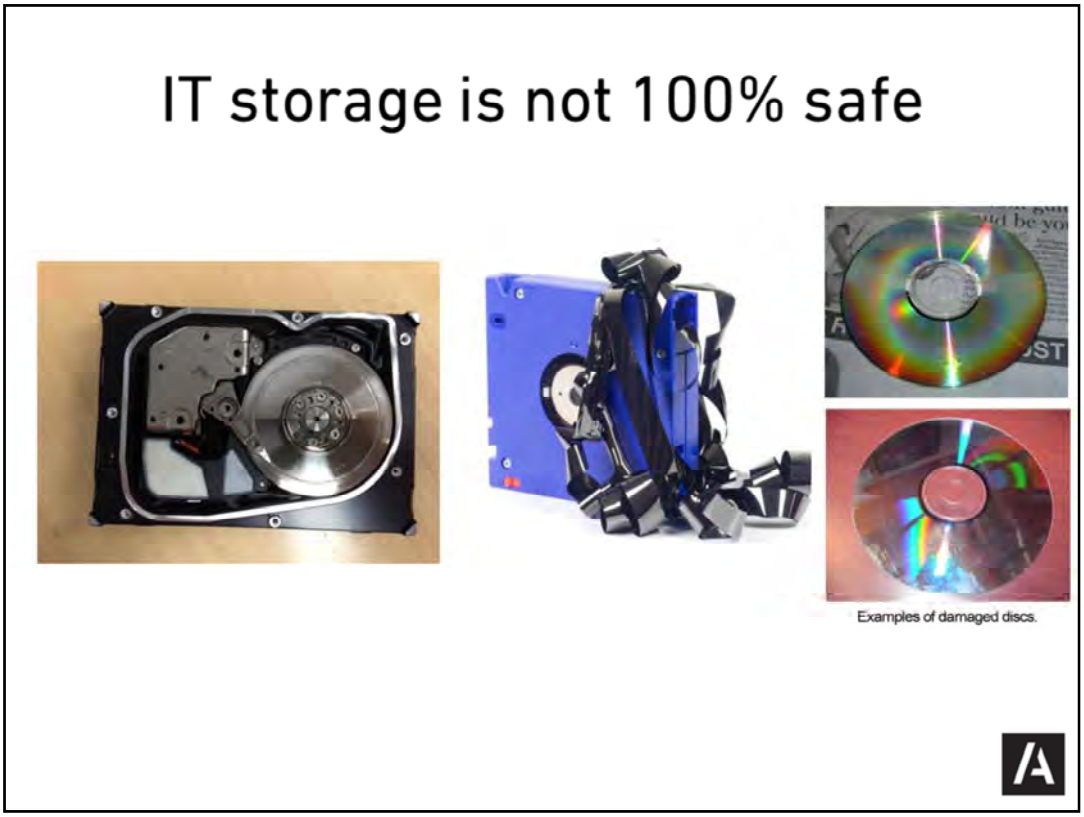
And there's no sign of it getting better

As humanity generates every more data, technological advances are made to store and access it.

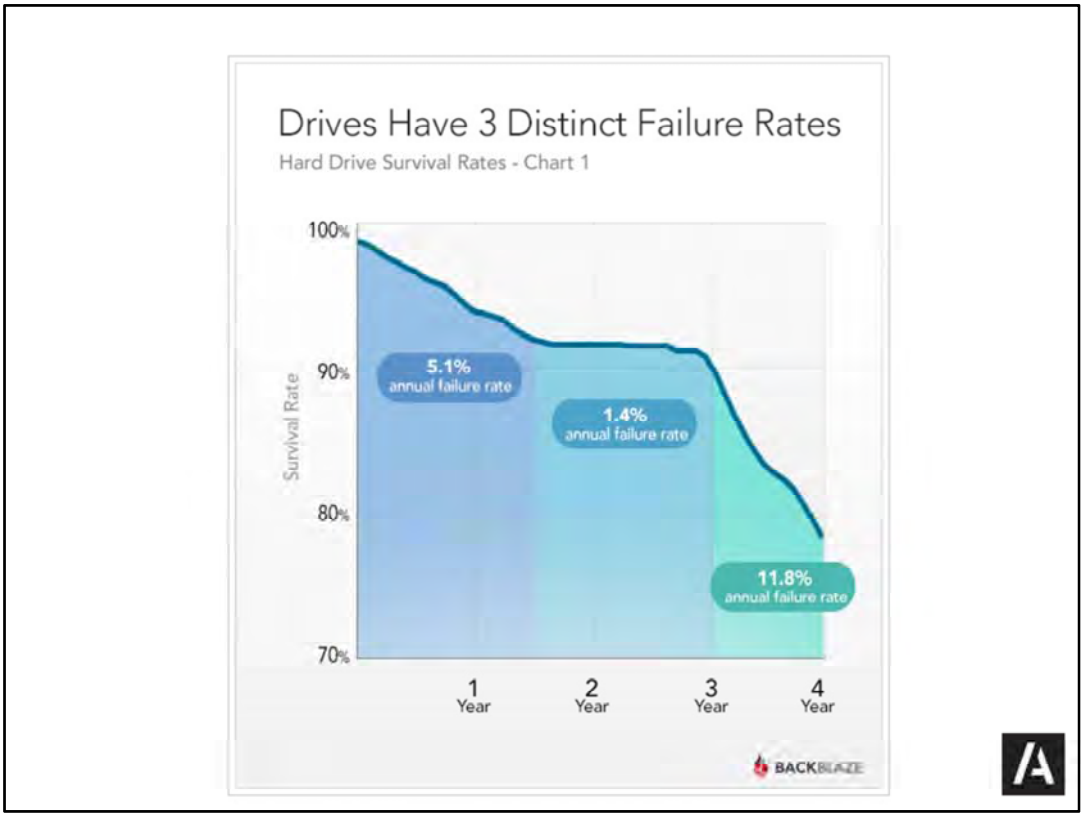
Advances get faster and storage gets cheaper, but the new technologies created last for less time – they become obsolete ever sooner.

Change happens faster.

You can try and 'beat the system' e.g. by using 'long lived' media and try to apply more traditional 'paper archiving' style processes. But you'll probably get caught out by no longer having the drives, software and systems to read that media – and that's assuming it really does last as long as the manufacturers claim – which it often doesn't.



Or you can use IT storage – tapes, drives, discs – but IT storage isn't 100% reliable so you need to take active steps to protect data and keep everything running as it should – no 'file and forget' as you might do in the paper world.

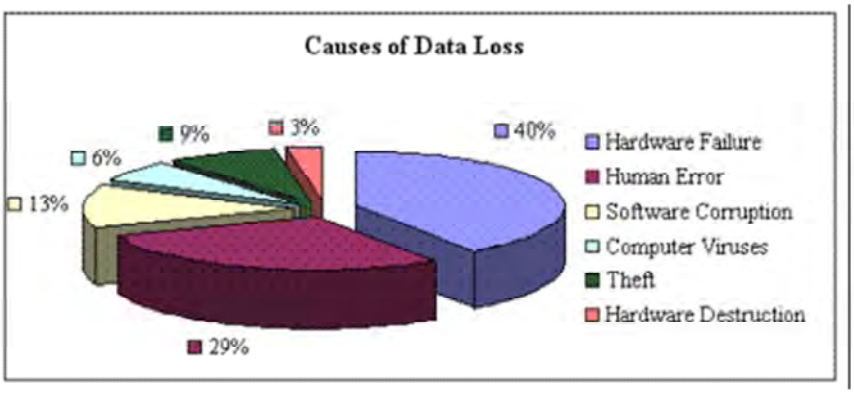


And just to illustrate this, BackBlaze have just released data on the failure rates they see for hard drives. They have 75PB of storage and have gathered stats from 25,000 hard drives. They confirm what others have already seen – hard drives fail. There is less than an 80% chance on average of a hard drive still working after 4 years in a storage server.

That's why lots of extra data protection technology is needed on the top – e.g. RAID arrays – and is a perfect example of why an archiving strategy based on storing hard drives instead of paper is doomed to failure.

http://blog.backblaze.com/2013/11/12/how-long-do-disk-drives-last/

People cause data loss too



Cause of data loss	Perception	Reality
Hardware or system problem	78%	56%
Human error	11%	26%
Software corruption or problem	7%	9%
Computer viruses	2%	4%
Disaster	1-2%	1-2%

But there are other ways to lose a lot more data that include system crashes, failures to backup, and human error. Indeed, human error can be a major factor – which doesn't mean deliberate loss – it means accidental loss through processes and procedures that aren't designed correctly or haven't been followed properly.

This is why you should keep multiple copies of data in multiple locations and ideally manage them by different sets of people.

http://www.zdnet.com/blog/storage/how-data-gets-lost/167
http://www.acspc.net/dataloss.php
http://www.thedatarecoveryblog.com/2013/03/27/backup-and-data-loss-survey-results/

There's plenty more ways to lose data!

- Technical obsolescence, e.g. formats and players
- Hardware failures, e.g. digital storage systems
- Loss of staff, e.g. skilled transfer operators
- Insufficient budget, e.g. digitisation too expensive
- Accidental loss, e.g. human error during QC
- Stakeholders, e.g. preservation no longer a priority
- Underestimation of resources or effort
- Fire, flood, meteors, aliens…

It is also why risk assessment is such an important tool as it allows the same approach to be used across a very wide range of ways in which data can be lost or compromised. Just considering long-term storage throws up a lot of risks – consider all those blue and yellow transitions in the slide earlier – and then add on the top the risks of not funding the setup properly or not having the right resources allocated.

So what can be done to help?

The first port of call is existing digital preservation standards and best practice.

The ones to look at first are those that take a risk assessment/audit based approach to building or using a Trusted Digital Repository – i.e. they define what should be present in any solution that aims to keep content safe, secure and accessible for the future.

ISO16363, which is the new 2012 ISO standard for Trusted Digital Repositories is a very good place to start and it builds directly on ISO27001 on information security. Drambora is notable as it takes a risk assessment approach. The Data Seal of Approval is also worth investigation as it takes a lighter touch to assessment and hence is easier to implement.

TDR and Drambora both explicitly address the risks associated with storage.

http://datasealofapproval.org/en/
http://www.repositoryaudit.eu/
http://www.iso16363.org/
http://public.ccsds.org/publications/archive/652x0m1.pdf

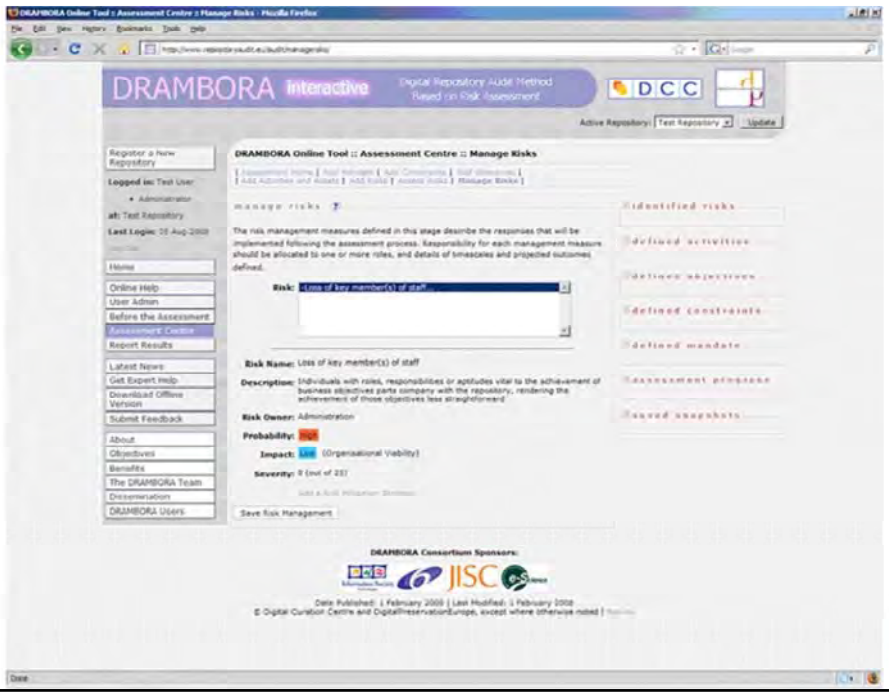
Others have trodden the road already

- Organisational risks
  - Understanding, governance, value
- Process risks
  - Business, IT and IM processes
- Operational risks
  - Technology failure, lock-in, obsolescence
- Continuity risks
  - Failure in integrity, availability, usability

The useful thing about existing digital preservation standards and techniques is that they've already been applied in practice by many organisations, especially large memory institutions, which means that there's supporting information too on how to apply them in anger. For example, the National Archives in the UK used a risk assessment approach, including indentifying some of the continuity risks I mentioned earlier.

http://www.bl.uk/blpac/pdf/decodingblake.pdf

DRAMBORA



Drambora has its own online toolkit for making risk assessment easy and this guides you through the process, so that's an easy way to get started.

Example risks

Risk ID	Title	Example
R30	Hardware Failure	A storage system corrupts files (bit rot) or loses data due to component failures (e.g. hard drives).
R31	Software Failure	A software upgrade to the system looses or corrupts the index used to locate files.
R32	Systems fail to meet archive needs	The system can't cope with the data volumes and the backups fail.
R33	Obsolescence of hardware or software	A manufacturer stops support for a tape drive and there is insufficient head life left in existing drives owned by the archive to allow migration
R34	Media degradation or obsolescence	The BluRay optical discs used to store XDCAM files develop data loss.
R35-R38	Security	Insufficient security measures allow unauthorised access that results undetected modification of files

Drambora comes with a set of pre-identified possible risks, many of them associated with storage of information, for example the risks shown here cover some of the risks of data corruption, system failures, technical obsolescence and security breaches associated with storage.

Prioritising risks

	Impact		
Likelihood	Major	Moderate	Minor
Likely			
Possible			
Unlikely			

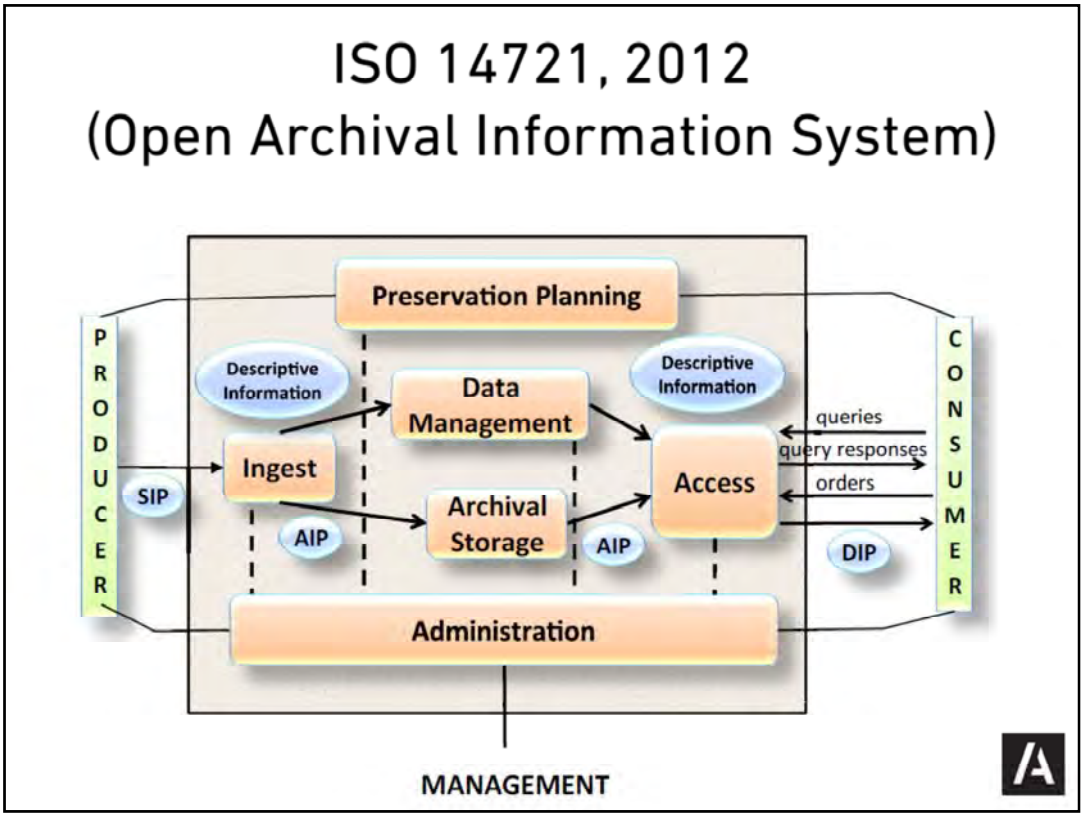
Having identified the risks, they should now be prioritised in terms of their impact, i.e. what would happen if the risk occurred, and how likely they are to occur. This allows resources to be focussed on the ones that are of most concern. This is all pretty standard risk management stuff and digital preservation is no exception.

ISO 16363, 2012
(Trusted Digital Repositories)

- Cause
  - e.g. failure to maintain storage systems
- What is affected
  - e.g. eTMF, audit trail, contracts
- Consequence
  - e.g. fines, loss of reputation, recreate content
- Priority
  - e.g. high
- Best practice
  - e.g. technology watch, migration planning
- Mitigate, avoid, accept, transfer
  - e.g. regular tests and migrations, multiple copies of data

Deciding what to do about the risks comes next, e.g. whether to mitigate them, find a way of avoiding them altogether, moving the risk to someone else, e.g. a supplier, or simply accepting the risk as something that has to be dealt with if it happens.

TDR and Drambora are both useful here as they provide best practice and example ways to mitigate risks – they don't just help you work out what the problem is, they help you fix it.



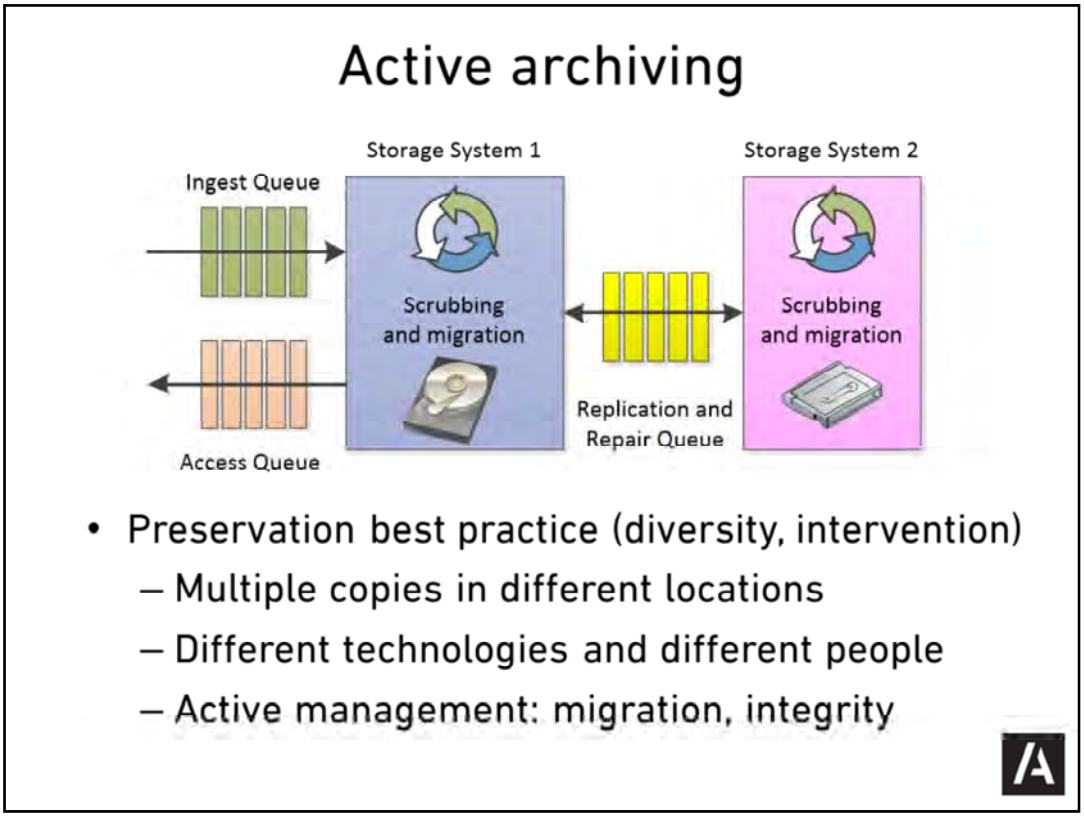
The other main standard worth mentioning is the Open Archive Information System – from the Consultative Committee for Space Data Systems, or other wise known as 'brought to you by NASA', although for the recent iteration many more space agencies have been involved, so it's a major initiative. These are the same guys who developed ISO16363 TDR

OAIS provides a framework for long-term retention and access to digital content using several key principles, for example:

- Ensuring you capture all data and metadata up-front according to a submission agreement to ensure you have everything needed in the long-term, which is important as it breaks the dependency on the original producers of the content to understand what it means and how to interpret it.
- The need to support what OAIS calls a Designated Community, which are the people who will need to use the content, for example regulators or internal reuse. This is important and the designated community defines what key characteristics need to be preserved in digital content.
- Then in the middle is the internal planning and managing of the storage and preservation actions so that the designated community can be served and the key characteristics of the content that are important to this community are maintained.

This approach is a bit different to Electronic Document and Records Management (EDRM) systems which, by and large, focus on the earlier part of the lifecycle where content is first created, not when it subsequently needs to be preserved. So what OAIS provides is a framework for assessing whether an EDRM has the right features to support digital preservation.

[1] http://public.ccsds.org/publications/archive/650x0m2.pdf



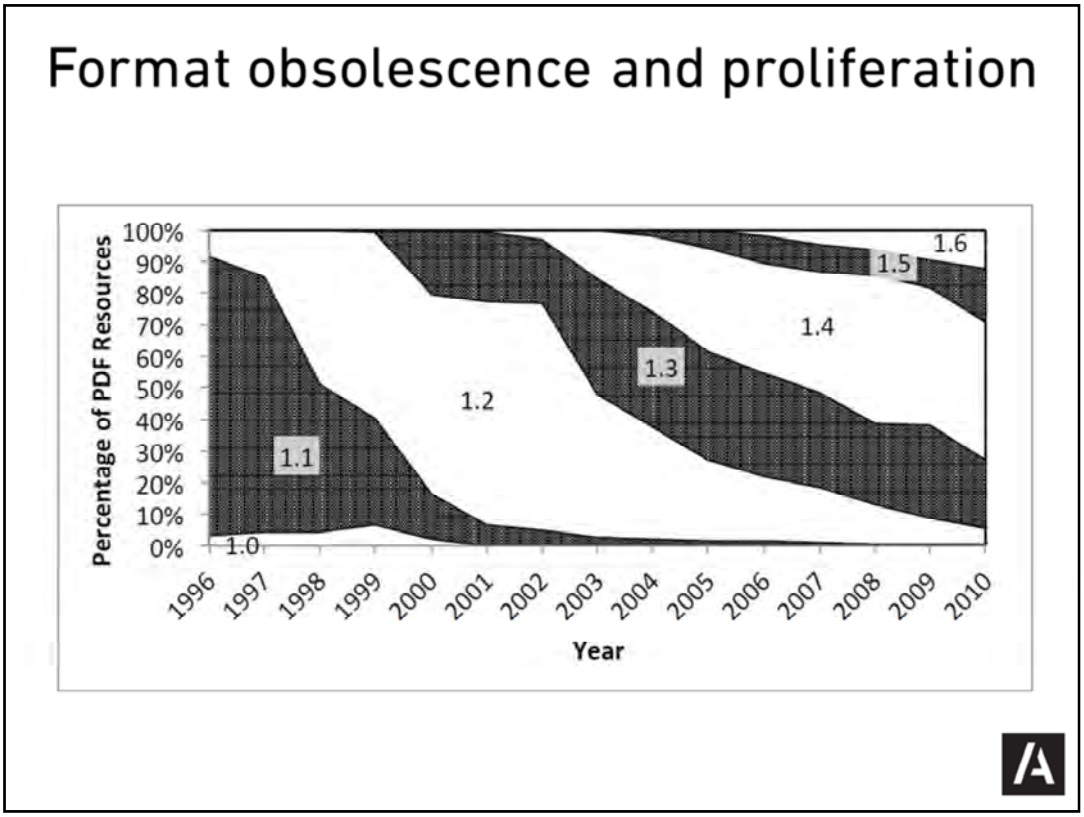
In the centre of OAIS is archival storage and this underpins a TDR too. The basis of this is pretty simple:

You make multiple copies of the data and you keep them in different locations.

You use diverse technologies to spread the risk of failures, which effectively means not all eggs in one basket

And you actively manage these copies by
(a)migrating to new storage or formats to address obsolescence, and
(b) by regularly checking and repairing any loss of data integrity (which is why having multiple copies is so important – if there is a problem with one of the copies then it can be replaced by replicating one of the other good copies). This is where checksums come in so the integrity of each file can be monitored, any corruption or loss detected, and a known good copy found to replace a corrupted copy.

These days there really is no excuse for losing digital data. But it does require the right skills and expertise, the right policies and procedures, and the right infrastructure. It's also not something you get out of the box from enterprise storage systems.



Next consider the applications that create and understand the data and formats used to store the data in.

This picture shows the use of different versions of PDF as seen by the British Library when harvesting web content in the UK [1] from 1996 to 2010.

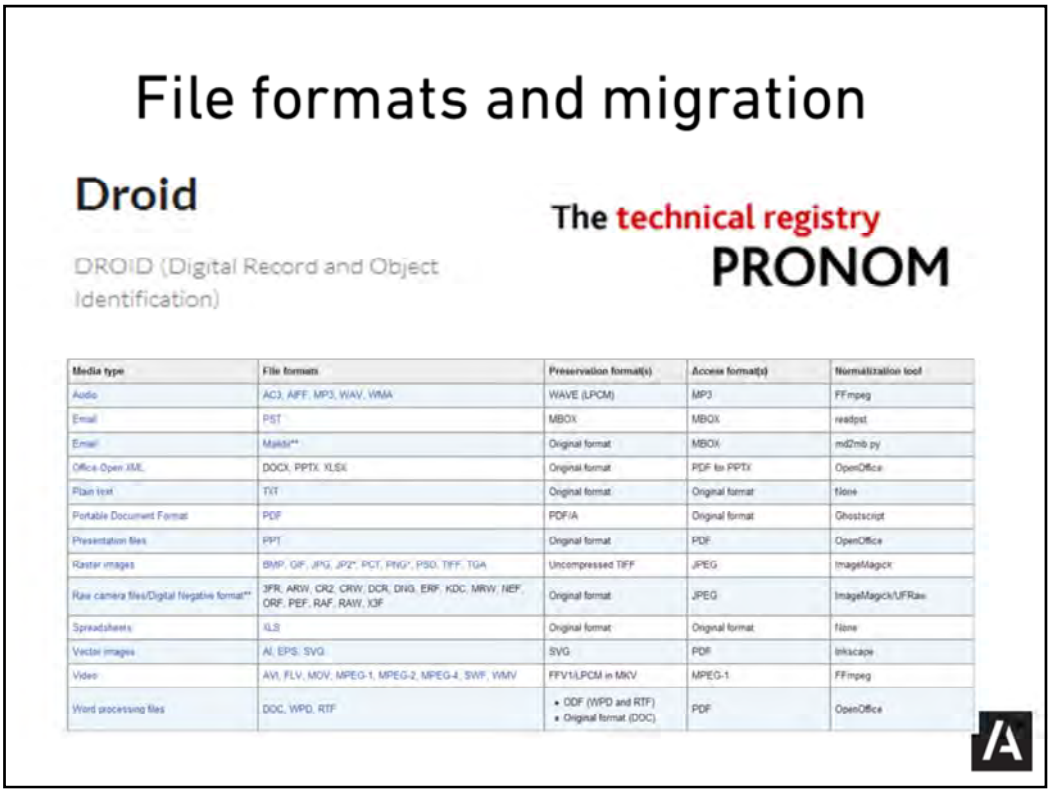
New versions of PDF are available every few years, they get widespread adoption, but then rapidly get superseded by new versions so their use tails off, although it does take a long time to die completely.

For PDF backwards compatibility is good so obsolescence isn't so much of a concern, especially with PDF-A, although it does highlight another issue which is the sheer number of different versions that may be in an archive at any one time.

But what about all the other types of content, for example correspondence, supporting data from laboratory systems, patient scans, etc? That's the stuff to really worry about.

There was an initiative set up November last year on file formats as a crowd-sourcing exercise to list file formats and their specifications. This already contains 140 formats for biological, biomedical and medical imaging data [2] I bet that's the tip of the iceberg. How many of those will be obsolete in the next 10 years?

[1] http://www.scape-project.eu/wp-content/uploads/2012/11/iPres2012_Formats-over-time.pdf
[2] http://fileformats.archiveteam.org/wiki/Scientific_Data_formats

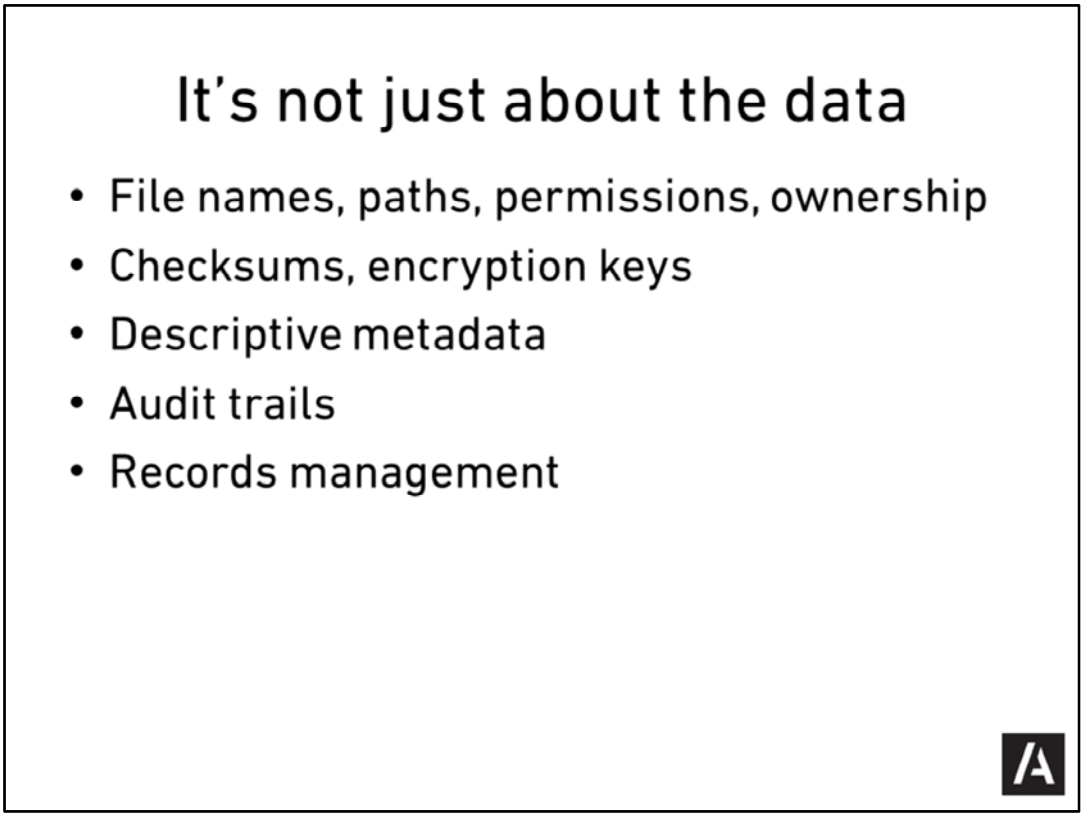


In the library and archive community this is a recognized problem. For example, there are format identification tools and services such as PRONOM and DROID from the national archive [1], and guidelines on what formats to choose for long-term accessibility. But support for scientific data formats is very limited.

So the most common strategy is migration to an open, well documented and long-lived format - often when the data is first archived, but sometimes at a later date if the format of the data isn't immediately of concern. This is often PDF-A for documents. But again the challenge is all that other supporting digital content.

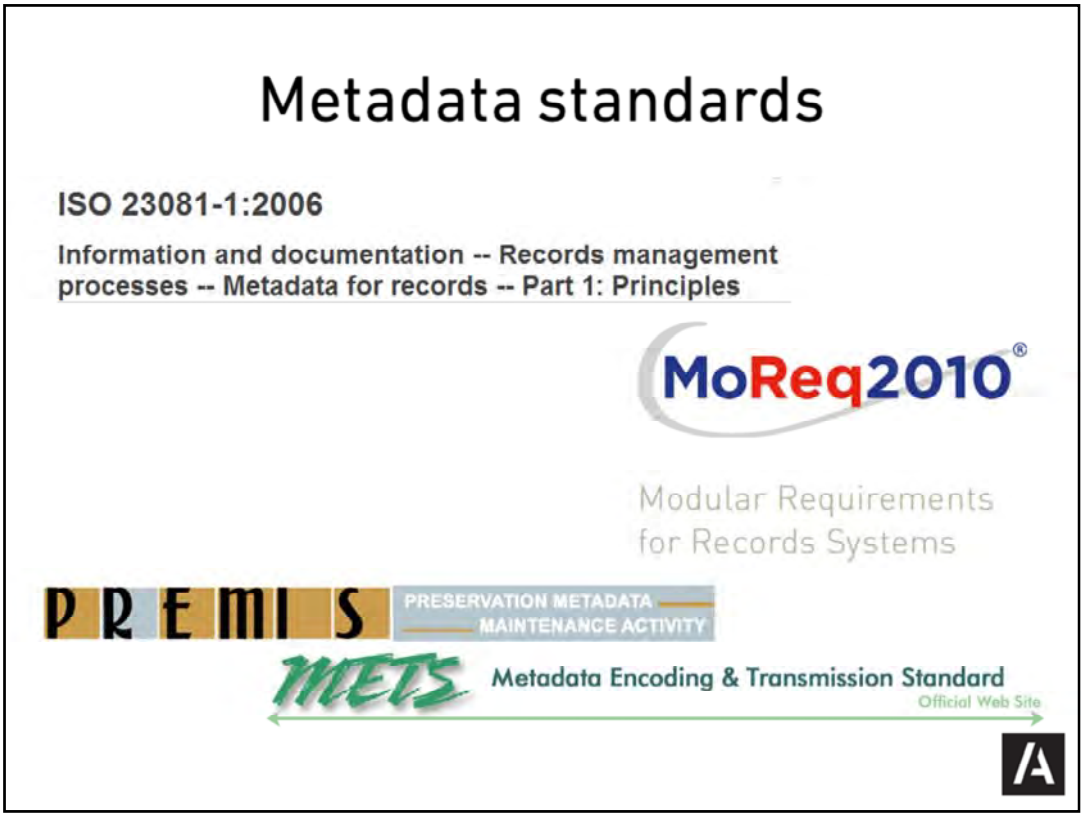
The approach is typically to migrate multiple data formats into one or more canonical forms, as shown in this table for the Archivematica [2] tool. The key thing is that the format for long-term preservation isn't typically the format data was first created in, and neither is it the one best suited for access and use. The objective is to think about longevity first and foremost.

[1] http://www.nationalarchives.gov.uk/information-management/our-services/dc-file-profiling-tool.htm
[2] https://www.archivematica.org/wiki/Main_Page



It is important to recognise that it isn't just the contents of data that needs protection from a data integrity and authenticity standpoint. Unless the metadata that describes the data, is safe, e.g. the original name/path/permissions for a file, then the ability to use the data may be lost. Likewise, unless audit trails, encryption keys or checksums are also kept safely and securely then the ability to assert authenticity, confidentiality and integrity can also be lost – which is particularly important if this forms part of evidential weight.

This means that the same risk-based approach taken when working out how to store data should also be applied to all the metadata as well. It's unlikely to be sufficient to simply export files from a document management or records management system but leave all the metadata behind that describes what that data is.



This is where preservation and records management standards come in that worry about metadata.

There are standards for records management [6], for example the ISO 23081 series, which addresses metadata for records management. And there's MoReq2010 [1] which is the latest in a series of requirements frameworks for records management systems. These build on ISO 15489, which is a records management standard that requires records to have the characteristics of authenticity, reliability, integrity, and usability.

But there are also standards for metadata from the digital library and digital preservation community, which include:

PREMIS, which is a standard for metadata associated with the preservation of digital objects, and can for example might contain checksums, signatures, format descriptions, file structures, and a list of events associated with preservation such as migrations.

And

METS, which is a standard for encoding various types of metadata into a form that can easily be stored or exchanged. Metadata put into a METS wrapper might include administrative metadata, e.g. PREMIS metadata, but also descriptive metadata of the content of a digital object, and how the components of an object are structured, e.g. a set of files.

These have a lot in common, e.g. they have the concept of events that need to be recorded which are part of the lifecycle of an object or record, e.g. when it was created, migrated, disposed of etc. And they have the concept of fixity, authenticity, usability etc. that align with regulatory requirements.

So, whilst there are:

- increasingly well developed standards for digital objects themselves, e.g. eTMF[5] or eCTD [4], and
- regulations that require the retention of these objects in a way that can assert integrity, confidentiality, authenticity and accessibility,
- what the metadata standards do is to allow the preservation of these objects to performed, managed and recorded a lot more effectively

And most importantly these standards allow it to be done in a vendor neutral way, i.e. they do not assume any specific products, tools or infrastructure. Indeed the opposite is true – preservation metadata can be structured and recorded using simple XML formats and stored in files on a file system.

[1] http://moreq2010.eu/
[2] http://www.loc.gov/standards/premis/index.html
[3] http://www.loc.gov/standards/mets/
[4] http://estri.ich.org/eCTD/eCTD_Specification_v3_2_2.pdf
[5] http://www.etmf.org/
[6] http://www.armaedfoundation.org/pdfs/V_Jones_RIMStandards_Update2012.pdf

So all of this probably sounds quite complex – there's a lot of standards and best practice to get to grips with.

The next section is how to get started with a simple strategy.

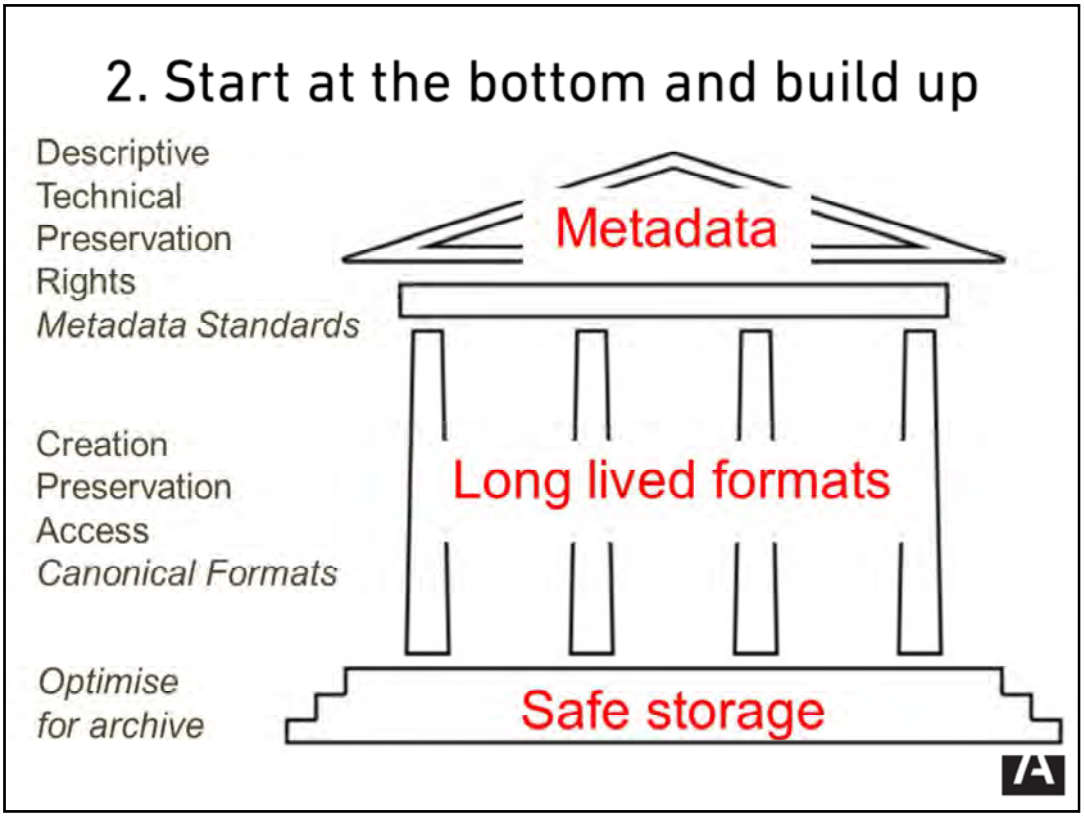
1. Assess and manage the risks

- ISO27001 Information Security Management
- ISO16363 Trusted Digital Repositories
- DSA Data Seal of Approval

Start with applying a risk management methodology – just as you would to any other part of the process.

Use the help out there for thinking in the right way and making sure the bases are covered, even if you don't necessarily expect to be certified against these standards. Take advantage of case studies and practical notes from large archives that have already gone down this path – even if they aren't from the lifesciences sector they will have similar objectives.

[1] http://www.27001-online.com/
[2] http://public.ccsds.org/publications/archive/652x0m1.pdf
[3] http://datasealofapproval.org/



Then tackle the digital preservation problem from the ground up.

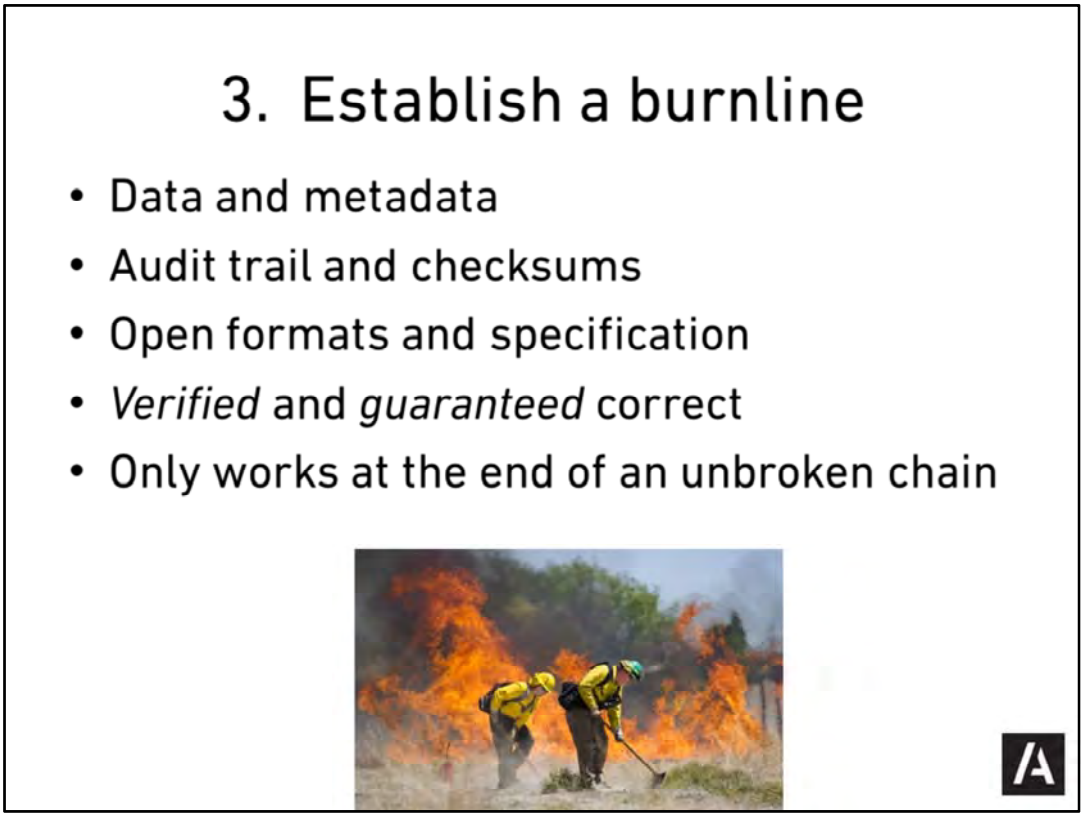
The key thing is solid foundations – storage and storage management that won't lose your data

Then comes file formats – the approach here is to choose a long-lived format if possible, e.g. PDF-A that is simple, open and easy to process.

Finally, on top of the house is the roof, which means metadata to describes what data is being held and allows it to be found again in the future. Metadata can generally be split into four types:

Descriptive metadata that says what's in the file, e.g. the eTMF XML backbone
Technical metadata that says what the format is, e.g. PDF
Preservation metadata that says what to do, or has been done, to the data to keep it accessible, e.g. migrations
Rights metadata that says who owns or can use the data, e.g. IPR

These should all be captured in an open format, e.g. XML.



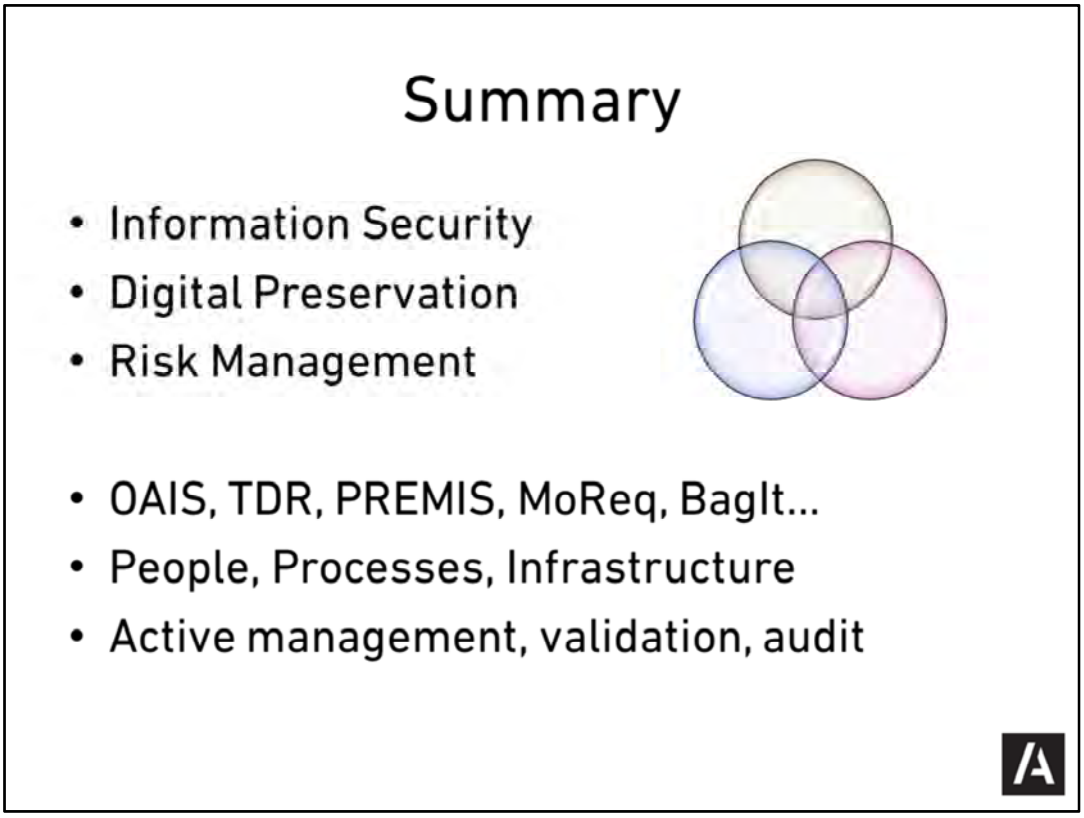
Then make sure that you have a complete and independent copy of all the data and associated metadata. This is your backstop, your fall-back, your get out of jail free card. This is what you rely on if everything else goes wrong.

Establish a 'burnline'. This is a term I've borrowed from Neil Jeffries at the Oxford Bodlean Library and refers to safeguarding research data. The idea if that if there's some form of disaster, the equivalent of a burning building, then the only thing that you need is the storage system used to hold the data. This is because on storage you put a complete record of all the data and associated metadata – this is in addition to, or instead of, using complex records management systems, databases, sharepoints or whatever else. This data and metadata is held in open formats and using open standards. Everything can be rebuilt from that if necessary.

This is just good preservation practice and provides extra assurance in a way that is:

- Simple and removes complexity.
- Removes dependencies, including lock-in to vendor systems.
- Lowers the risks

And if everything else is in place, then you'll be able to assert that the burnline copy of the data is complete and correct.



So just to summarise so far before spending the last 5 minutes looking at how Arkivum can help. Hopefully I've shown that in order to store data for long periods of time and in a way that will meet regulatory requirements then you need to apply three main techniques:

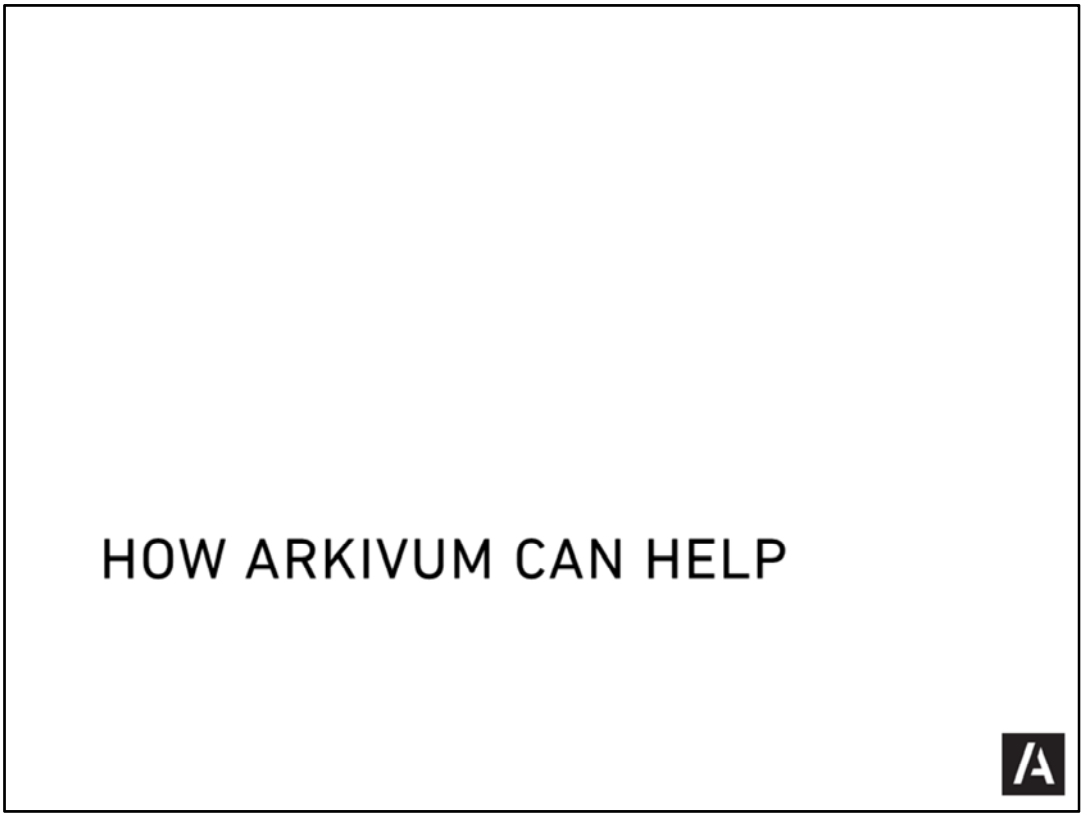
- Information security
- Digital preservation
- Risk management

You can do this by taking advantage of the widely available standards and best practice in each of these areas.

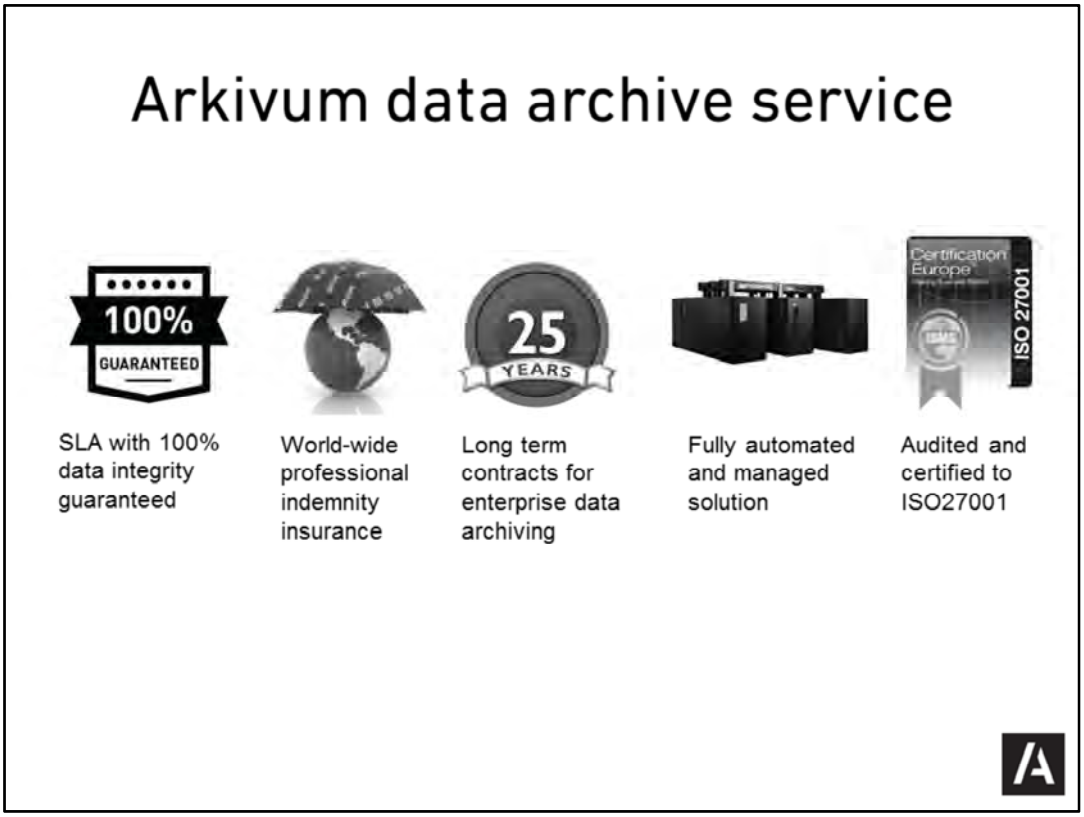
This allows you to tackle the risks to long-term data retention in a managed way.

- Key to doing this is recognising that it isn't just a technical problem but requires the right people and processes to be in place
- This in turn means you need the right skills and infrastructure
- Which can then be applied in a proactive way to keep data alive
- There is no 'file and forget' in the digital world – data needs active steps to be preserved
- And these active steps all require validation and audit trails so regulatory requirements can be met.

You might have expected me to talk more about specific types of storage or technologies, but storage technologies will come and go – what matters is an approach to the problem that is technology independent and can be applied in a way that recognises change rather than tries to avoid it.



So now for the last few minutes of how Arkivum fits into all of this.



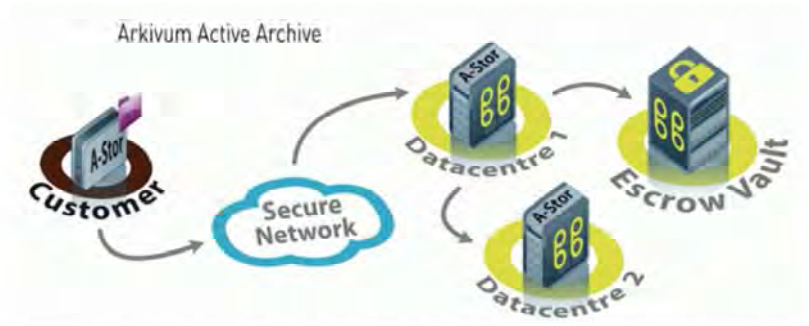
Arkivum provides online data archiving as a service for those organisations that need to keep data for the long-term for compliance or reuse. We have customers in healthcare, life sciences, energy and construction to name but a few.

The company was founded as a spin out of the University of Southampton and is based on expertise staff have in from working with large archives over the last 10 years.

The key features of the service are that:

- We guarantee the integrity of data in the service as part of the contract
- The guarantee is underpinned by insurance that covers costs incurred by customers if there were to be a data loss
- Long-term contracts are available with PAYG and Paid-Up payment models
- The solution is fully automated and managed which takes away the complexity of having to do it yourself
- We've been audited and certified to ISO27001 giving assurance that we can help you meet regulatory requirements

Arkivum approach



- Best practice: 3 copies, 3 locations, online and offline
- Active archiving: integrity, obsolescence
- Risk management: ISO27001, Drambora, TDR

The approach we take is to follow data preservation best practice. Three copies of customer data are held in three known UK locations. We use checksums to actively manage data integrity through regular checks and we do regular media and infrastructure migrations to counter obsolescence and to ensure costs remain low. All data is encrypted for security when it is still within the customer network and only then is it uploaded to the Arkivum service. The solution is provided as a service, but can also be installed and operated in-house if required.

Basically, we take care of that 'blue and yellow stuff' in that diagram I showed earlier. And it's our skilled and dedicated staff that do this that you're also getting access to, not just the infrastructure we use.

We are certified to ISO27001,. We follow the OAIS model. We have done our own DRAMBORA risk assessment and we apply ISO16363 TDR principles. What we've done is all the hard work on following digital preservation standards and this takes much of the burden off our customers.

Finally, one of those copies of customer data is held at a third-party site with a three way agreement in place with us, the third-party and our customers so that if they want to leave our service, or if we can no longer provide the service we agreed, then the customer gets direct access to a complete copy of their data on media that they can take away.

Ingredients

- Skilled and trained people
- Validated processes and procedures
- Comprehensive risk management
- Specialist infrastructure
- Economies of scale
- Independent audit and validation

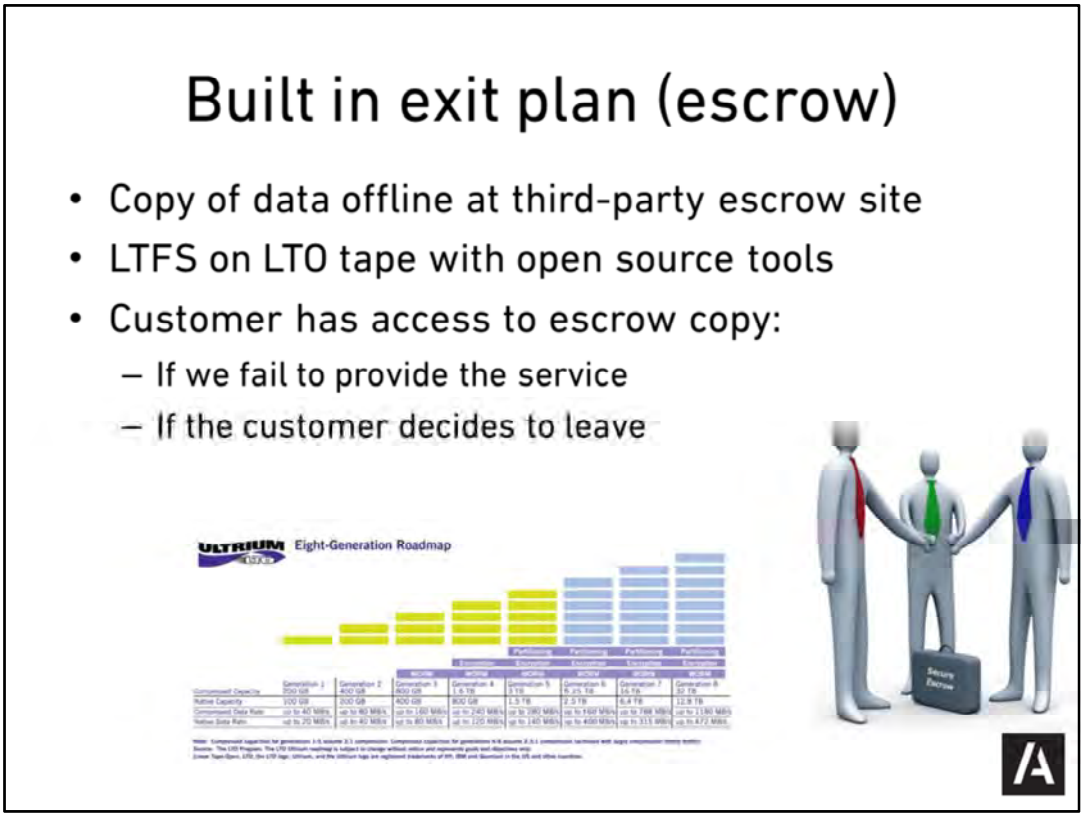
And as I mentioned before, the role of having the right people and procedures in place cannot be underestimated.

Customers of our service are buying into:

Skilled and trained people that follow validated processes and procedures and do this within a framework for comprehensive risk management.

Specialist infrastructure needed for archiving and long-term retention of data, where we keep costs down and data safety high through major economies of scale.

Independent audit and validation, e.g. ISO27001 but also due diligence by our customer base, which all provides assurance that everything is being done properly.



Much of our service is built on modern data tape, known as LTO, which is a reliable and industry standard way to store large volumes of data. Because we use data tape, we can create an offline copy of customer data that is lodged with a third-party under a three-way agreement between us, them and our customers. The use of LTO along with an open way of storing files on tape called LTFS and a set of open source tools means restoring data from escrow is easy and has no lock-in to hardware or software vendors, including us. Customers can access the escrow copy of their data if we either fail to provide our service or if the customer decides to leave. This gives customer reassurance and an easy exit-strategy if the need it, which is something you don't see with cloud storage providers.

Because we provide an end-to-end chain of custody that ensures data integrity, authenticity and confidentiality between our customer's storage systems and our storage service, the data and metadata on the escrow tapes is a ready made burnline. As someone from a large archive said to me, think about how you will get out, i.e. your exit plan, before you think how you will get in to a preservation service or solution. Our escrow model and burnline provides exactly that.

Finally, if you are concerned about letting your content offsite, then everything I've described so far can be deployed on a customer site and managed there with the same contractual guarantee of data safety and insurance backing.

http://www.lto.org/
http://www.lto.org/technology/ltfs.html

www.arkivum.com
Matthew.addis@arkivum.com

QUESTIONS