


100% data integrity guarantee
Keep your data safe & secure forever

How to Keep Genomics Data Safe, Secure and Accessible for the Long-Term

Dr Matthew Addis
Arkivum Ltd
NGS Data Congress London 2013

© Arkivum Ltd 2013



Contents

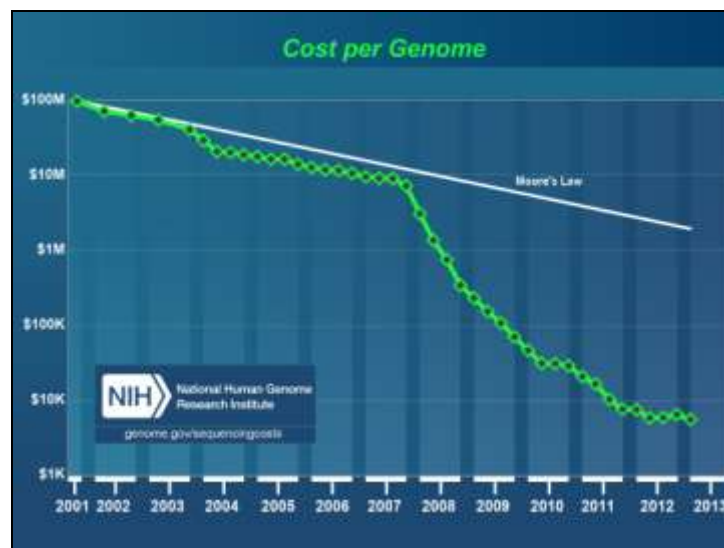
- Just keep the sample and re-sequence?
- Keeping digital data: so what's so hard about that?
- What are the practical solutions?

“Security is of key importance to our business, Arkivum’s A-Stor Pharma service allows us to store our encrypted data for the long term in a cost efficient way that is entirely scalable and reduces pressure on our internal IT infrastructure.”

Dan Watkins, Oxford Fertility Clinic

WHY NOT RE-SEQUENCE?

Slide 4



The cost of sequencing has fallen by a factor of 1000 in the last five years. Sequencing is a few thousand \$ per genome. The rate of improvement may not continue at the current rate, but it's certainly outstripping the rate at which storage costs are falling.


This falling cost and commoditisation of gene sequencing are driving an increase in data of 100% CAGR.

So if the cost of sequencing is falling faster than the cost of storing data that is created by sequencing, then are we at the point where it's cheaper just to keep the sample and sequence it again instead of storing all the data?

<http://www.genome.gov/sequencingcosts/>

Data volumes

- Genome
 - 3Gb, 30x coverage = 200 GBytes
- Exome
 - 110 million reads, 100bp per read = 15 GBytes



© Arkivum Ltd 2013

The data volumes are large, especially for a whole genome, although it is still the case that most sequencing is done at the exome level or for individual genes.

Data may also be stored in multiple formats, e.g. the raw reads in FASTQ, alignment to a genome in SAM or BAM, and then SNP annotations in say VCF.

It all adds up, especially in projects like the 1000 genomes where the objective is to look for differences across a population.

<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2836519/>

<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3042601/>

<http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/>

Cost comparison (Human Genome FASTQ)

- Digital storage
 - £70 per year for storage and all retrievals
 - Less than 5 minutes to access data
- Re-sequence
 - £4000 (NHGRI, Jan 2013)
 - Days or weeks to regenerate data
 - Plus storage of the sample
 - Cost increases if no reference genome
 - Opportunity cost of sequencer and staff

© Arkivum Ltd 2013

ARKIVUM

Suppose we take the case of a human genome (3Gb) using FASTQ, i.e. the raw reads and quality scores that come from the sequencing. That's approx. 200GB of data at 30x coverage.

At Arkivum's current list prices, that would be approx. £70 per year to store. The cost of storage is falling, so the 20 year retention of that 200GB of data in the Arkivum service can be purchased for an up-front cost of £500.

That's still way lower than just the cost of re-sequencing the sample.

Re-sequencing isn't the only cost. The sample has to be stored and the resources used for sequencing are diverted away from other activities so there's an opportunity cost too. The cost increases further if there's no reference genome, i.e. de novo sequencing.

But it's the time it takes that's the real problem. Retrieval from data storage is measured in seconds. Re-sequencing is measured in days or weeks.

<http://www.genome.gov/sequencingcosts/>

Benefits of keeping digital data

- Reuse
- Regulation and compliance
- Intellectual Property
- Save money
- Better use of in-house resources

© Arkivum Ltd 2013

So we are for some time to come in a world where storing digital data from sequencing makes economic sense. But this isn't the only reason for keeping data and not just samples.

There are several reasons to archive data, which include:

Because you have to, for example meeting legal regulations on clinical trial data.

Or 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. realignment against a different reference genome

Or because it saves money as we've already discussed.

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.

GxP Regulations and Guidelines

- FDA 21 CFR Part 11
- EudraLex Volume 4 Annex 11
- OECD series on GLP
- MHRA GCP guide



© Arkivum Ltd 2013

It's worth looking a bit more deeply into the regulatory requirements, not because I want to scare you into keeping genomics data for compliance reasons, but because the regulations can actually be instructive on how to think about the problem of long-term retention.

In the GMP space there's 21 CFR part 11 in the US and Volume 4 Annex 11 in the EU. These are just some of the GxP regulations and guidelines that apply, for example OECD deals with GLP and explicitly talks about archiving.

http://ec.europa.eu/health/files/eudralex/vol-4/annex11_01-2011_en.pdf

<http://www.oecd.org/chemicalsafety/testing/oecdseriesonprinciplesofgoodlaboratorypracticinglpandcomplianceandmonitoring.htm>

[http://search.oecd.org/officialdocuments/displaydocumentpdf/?doclanguage=en&cote=env/jm/mono\(2007\)10](http://search.oecd.org/officialdocuments/displaydocumentpdf/?doclanguage=en&cote=env/jm/mono(2007)10)

Medical records



The Royal College of Pathologists
Pathology: the science behind the cure

The retention and storage of pathological records

DPA, FOI, Public Records Act, BS10008 and BIP0008, IG Toolkit, Retention Periods (Part 2), Audit Trails (Annex D3)

Records Management
NHS Code of Practice
Part 1

© Arkivum Ltd 2013

Regulatory requirements/guidelines can apply for almost all applications, from NGS as part of medical records when used as part of diagnostics or to identify treatments

<http://www.rcpath.org/Resources/RCPPath/Migrated%20Resources/Documents/G/g031retentionstorageaugust09.pdf>



And upstream when doing research.

Academic research is in a sense regulated, for example, the Medical Research Council publishes guidelines on Good Research Practice [1], which includes the need to retain research data and have a data management plan. This is about repeatable and verifiable science. And with the added driver of sharing the results of that science in an open way where possible to foster independent validation and reuse.

[1]

http://www.mrc.ac.uk/Ourresearch/Ethicsresearchguidance/Researchpractice/principles_guidelines/index.htm

Common Requirements

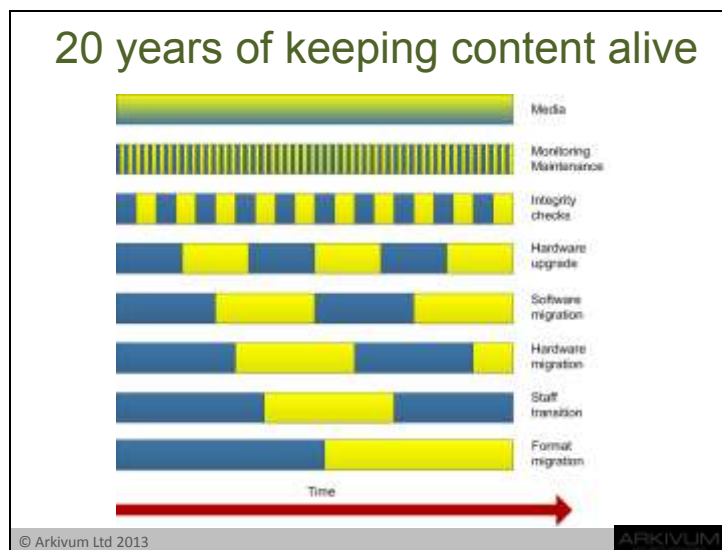
- Integrity
- Confidentiality
- Usability / ready access / readability
- Responsibility
- Risk management

© Arkivum Ltd 2013

But what all these regulations and guidelines have in common is not any specific recommendations on how to do long-term retention or how to provide future access, but instead some key characteristics that need to be achieved. In other words, they specify the outcome and not the mechanisms – which makes them a useful framework.



So if the regulations and guidelines provide a framework on what to achieve, then what's so hard about that?



These are just some of the things that will happen over 20 years of trying to retain data.

In the diagram, a change from blue to yellow is when something happens that has to be managed. In a growing archive, adding or replacing media, e.g. tapes or discs, can be a daily process, so is effectively continual. 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 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 key point is that long-term archiving is an active process and there's always some form of change going on. And when change happens there's always a risk that something goes wrong, and there's always the need to validate that the change has been effected properly. This all requires time, expertise and money. 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.

Digital Preservation

*"Digital information lasts forever -
or five years, whichever comes first."*

Jeff Rothenberg

© Arkivum Ltd 2013

ARKIVUM

The problem is one of Digital Preservation [1], which I'll come back to later, but for now it's summed up very nicely by Jeff Rothenberg from the Rand Scientific Corporate, who says that the natural lifetime of digital information is just 5 years [2]

If you want content to live longer then you need to make active steps to keep it alive.

[1] <http://www.dpconline.org/advice/preservationhandbook/introduction/definietions-and-concepts?q=definitions>

[2] <http://www.clir.org/pubs/archives/ensuring.pdf>

Perpetual Change

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

© Arkivum Ltd 2013

And keeping stuff alive means continual intervention and continual change.

It's about those blue to yellow transitions. Each one is a change. And:

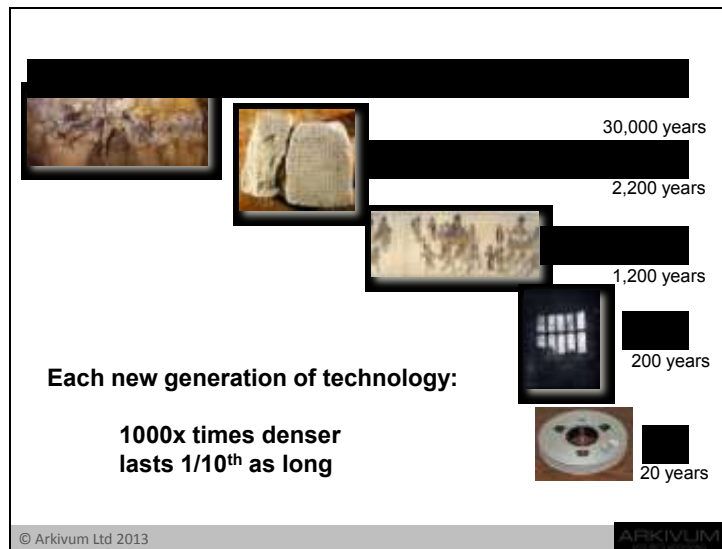
Change costs money

Change takes time

Change introduces risk

Change requires validation

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.



And if you think special long-lived storage is the answer, then ask a guy called Andrew Brown who was in the news in November last year. He asked his local hospital for a copy of an echo cardiogram that was performed on him in 2004. The BBC [1] reported that the Worcestershire Acute Hospitals NHS Trust said it would cost £2000 for him to be given a copy of his data. The Register reported that the trust said this "was not a cost-effective use of public money" [2]. The hospital does have his echo cardiogram data, which is stored on Magneto-Optical disk, but the hospital no longer has a drive that can read these disks as they have subsequently installed a new archive system [3]. Their supplier apparently said that they didn't stock the drive anymore as it was no longer in production and they would have to ship a drive in from the United States if the Hospital wanted to read the data. That's technical obsolescence in action – and in just over 6 years. This sort of thing really does happen. It illustrates why you need to think carefully whether 'long-lived' media is the right way forward. It's tempting to think that using specialist storage media that's designed to last for decades or centuries is the answer, including magneto-optical disks and other forms of 'archival grade' storage. But this technology often addresses a niche market, which means it can be harder for the companies who develop and sell it to make a sustainable business. The storage media might last for decades, but the companies who make it might not, or they are forced to move on to develop something new.

[1]<http://www.bbc.co.uk/news/uk-england-hereford-worcester-20235193>

[2]http://www.theregister.co.uk/2012/11/08/nhs_scan_2k/

[3]<http://www.whatdotheyknow.com/request/94658/response/237590/attach/2/attachme nt.pdf>



No storage is 100% reliable either, so there's another set of risks to deal with alongside rapid obsolescence. Hard drives are an example of supreme engineering, but they do go wrong – as a colleague in the storage industry said 'they are designed to be on the edge of not working'. 1% of hard drives simply fail each year – they won't work at all – and those that do work can suffer data corruption issues – even if used in storage systems (e.g. RAID arrays) that are meant to protect against data loss.

If you have data on USB drives on a shelf then worry.

And if you have data on a server, then also worry – but this time about the 10s of thousands of lines of software and firmware code that's in those servers and the bugs that we all know software contains.

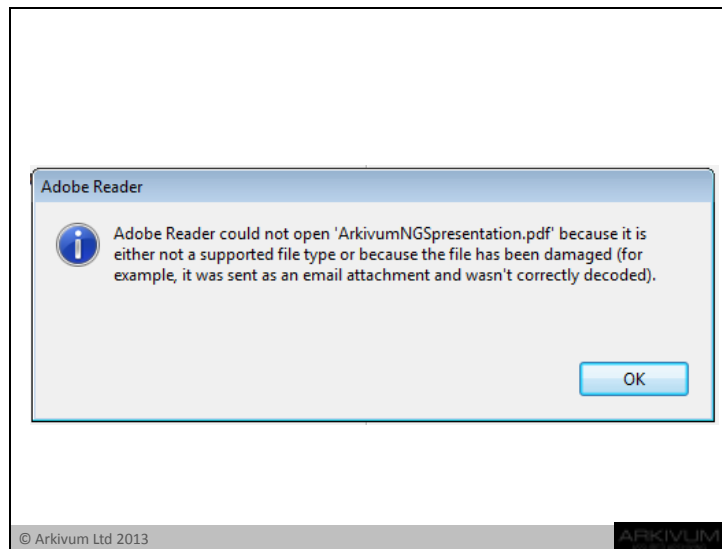
No storage media is 100% safe

- SSD
- CD, DVD
- Data Tape




Failure	Seen?	Devices exhibiting that failure
Bit Corruption	Y	SSD#11, SSD#12, SSD#15
Flying Writes	N	
Short Writes	Y	SSD#5, SSD#14, SSD#15
Unserializable Writes	Y	SSD#3, SSD#4, SSD#7, SSD#8, SSD#9, SSD#11, SSD#12, SSD#13, HDD#1
Metadata Corruption	Y	SSD#1
Dead Device	Y	SSD#1
None	Y	SSD#6, SSD#10, HDD#2

Other forms of storage aren't immune either – be it data tape, solid state or CD and DVDs.



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

It's typically 'all or nothing' when it comes to reading a damaged file. A failure often means heroic (and costly) measures to get the data back – if you can at all.

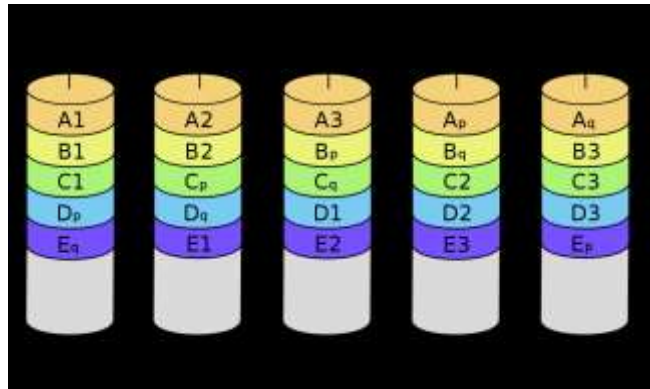
CERN did a test on their storage systems – which had been designed to keep data safe - 33700 files were written to storage and read-back again. (~8.7 TB). 22 were corrupted – and worse still, silently corrupted, i.e. they only knew because they checked each one.

CERN also did a test to see the impact of corruption. They took 10000 compressed files (zip) - 99.8 % wouldn't open if there was just a SINGLE bit error.

Slide 21



The IT Industry knows this already



© Arkivum Ltd 2013

ARKIVUM

This is no surprise to the storage industry, which is why storage vendors continue to develop ever more sophisticated ways to deal with failures and data corruption, for example RAID arrays and more recently self-healing file systems and other fun things.

Systems bring their own problems

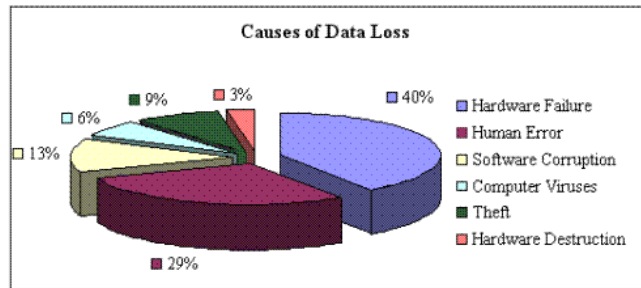
“Disk failures are not always a dominant factor of storage subsystem failures, and a reliability study for storage subsystems cannot only focus on disk failures. Resilient mechanisms should target all failure types”

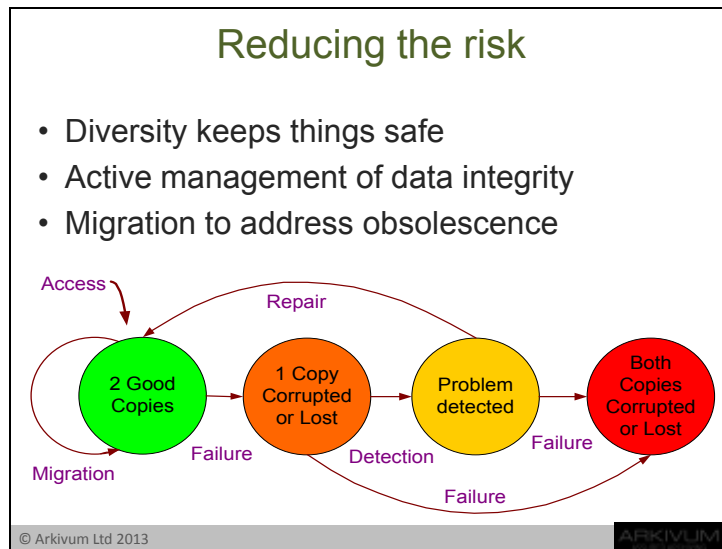
NetApp study of 1.8M HDD in 155,000 systems

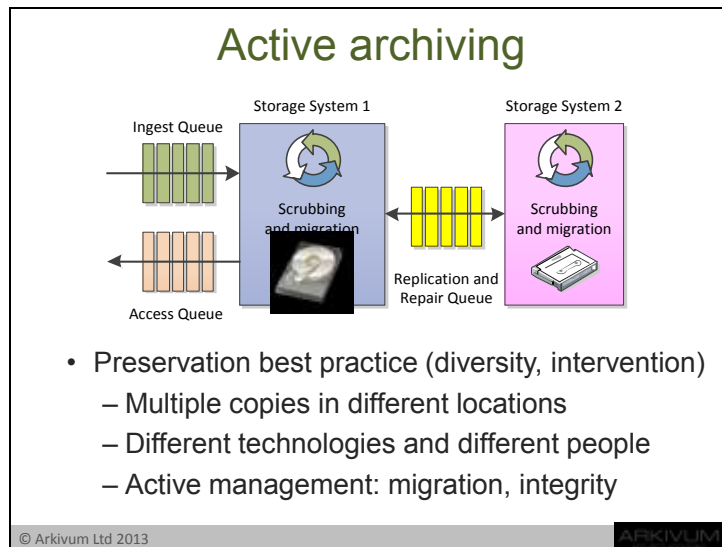
But these systems aren't foolproof – there can be many tens of thousands of lines of code in controllers and file system software – and that means bugs – and those bugs can mean data loss.

Netapp did a big study of nearly 2 million hard drives out in the field in a range of storage servers and found that there failures at all levels and worrying about disk reliability wasn't the place to start.

People cause data loss too!





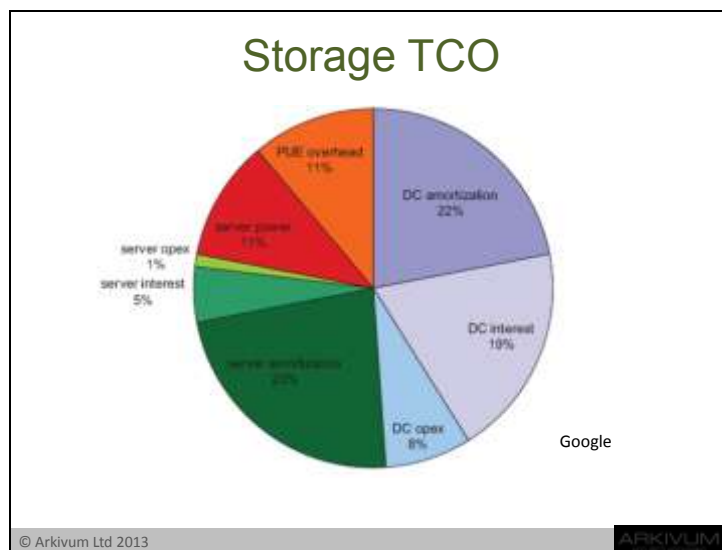


So this is what you do to get started at the safe storage level.

- 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)



So why not make more copies? Storage is free, right? This is a magazine advert from 1980. A 10 MB hard drive for \$3k. Today you can get 10 TB for close to \$300. That's a 10 million increase in capacity for the same cost.



But when looking at the costs of storage, it's important to recognise that servers and media are a small fraction. Power, space, cooling, maintenance all add up. That said, the Total Cost of Ownership of storage is still falling and in practice halves every 2-3 years.

Sustainable storage

- >100% CAGR in data volumes
- 20% CAGR in HDD capacity per £
- 2% CAGR in IT Budgets

IHS iSuppli
ComputerEconomics

- Store less
- Compress
- Better archiving

© Arkivum Ltd 2013

Which all means something has to give. The data volumes being created are outstripping the budgets available to store it.

There are three options.

- Store less. But this means an upfront cost of having to actively select what data to keep and the risk that important data is discarded.
- Compress the data. This can be well worth doing and even a factor of 2 compression makes a lot of difference for many TBs of data. But it's also not that easy to achieve high compression ratios, primarily because of things like quality scores (e.g. 60 characters per base pair) and errors in sequencing which mean that the full redundancy of the underlying base pair information can't be exploited.
- Find a way to archive infrequently accessed data at a much lower cost.

<http://ivory.idyll.org/blog/sequence-squeezing.html>
<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3083090/>

Cost v.s. safety v.s. access			
	Data tape on shelves	HDD in servers	Storage as a Service
Storage Cost	Low (media, shelves, climate control)	High (servers, power, cooling, maintenance)	High (fully managed service)
Access Cost	High (people retrieve and load media)	Low (internal network, automated)	High (bandwidth, charges for i/o)
Latent Failures	Low (data tape is reliable)	Med (‘bit rot’)	Low (replication and monitoring)
Access Failures	Medium (people handle tapes)	Low/Medium (depends on system)	Low (automated checks)
© Arkivum Ltd 2013			

In the end, you face a trade-off between cost, data safety and ease of access. There is no solution that has low cost, high safety, and instant access. The right approach is to pick a combination that gives the right balance.

Slide 31

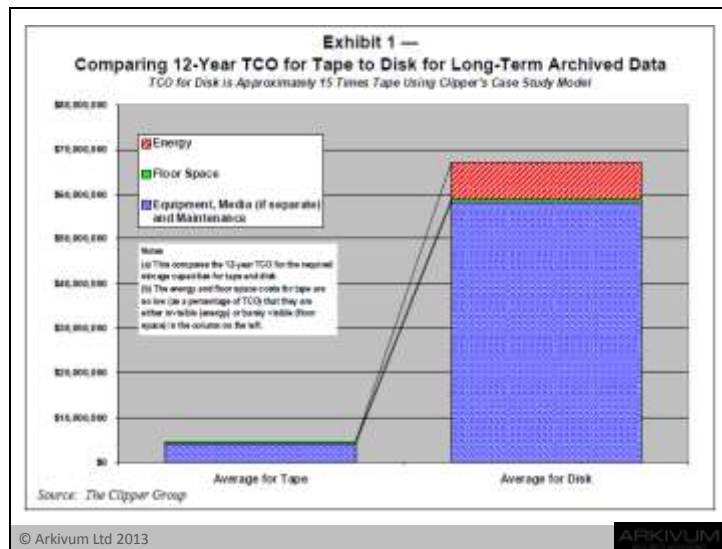


An extreme example is DNA itself as a digital storage media. This photo is from a seed reportedly germinated after it spent 32,000 in the Siberian permafrost. DNA can certainly provide robust long-term storage.

DNA Storage?		
1GB of data	DNA Storage	Hard Drive
Media life	10,000 years	10 years
Write time	5 years	10 sec
Read time	1 month	10 sec
Cost	\$1.2M	\$1

© Arkivum Ltd 2013

But whilst DNA for digital storage can be a very safe media, the cost and the ease of access is really really high. To illustrate this point, it's fun to look at the work Nick Goldman published in Nature in January on DNA as a storage medium.




The costs depend very much on the storage technology used. For example, the Clipper Group calculated that the cost of energy alone to power a hard disk storage solution over 12 years was more than the total cost of a data tape solution.



So whilst digital preservation standards and frameworks give you the tools you need, how do you get started.

1. Assess and manage the risks

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



© Arkivum Ltd 2013

Stage 1

Adopt a risk management approach. This helps you align with the regulatory requirements. It also allows you to apply existing and well developed and thought through standards for information security and trusted repositories.

ISO27001 is information security management and covers the integrity, availability and confidentiality of assets.

ISO16363 is less well known, but is a standard for trusted digital repositories that builds on ISO27001 and also uses a risk assessment model.

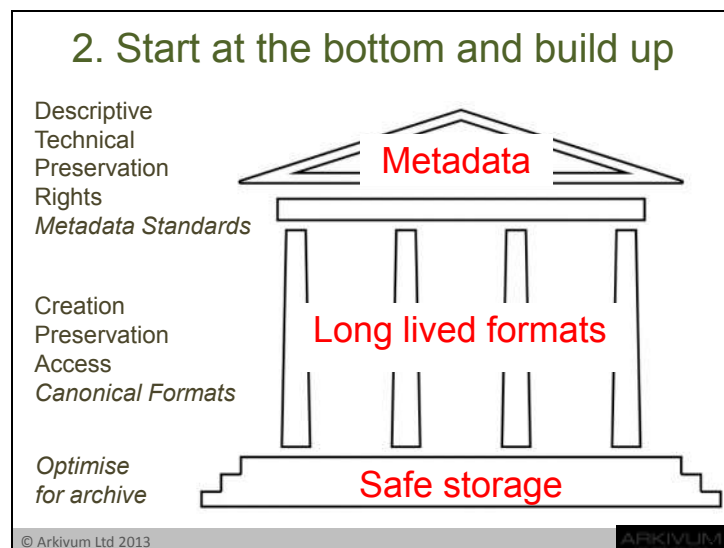
Then there's a more lightweight option called the Data Seal of Approval that doesn't require external auditing.

So there's 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.

[1] <http://www.27001-online.com/>

[2] <http://public.ccsds.org/publications/archive/652x0m1.pdf>

[3] <http://datasealofapproval.org/>



It's easy to get bogged down in complex solutions that try to crack all the problems at once.

So Stage 2 is a simple alternative approach of building a 'preservation house'. Start at the bottom and build 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. FASTA or 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. a particular sequence

Technical metadata that says what the format is, e.g. FASTA


Preservation metadata that says what to do, or has been done, to the data to keep it accessible, e.g. migration

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, ideally using a standard, e.g. METS and PREMIS.

3. Establish a burnline

- Metadata and data on a file system
- Open standards and formats
- Drive down costs and risks



© Arkivum Ltd 2013

Then finally Stage 3 is to support what's called a 'burnline'.

This is a term I've borrowed from Neil Jeffries at the Bodlean Library and refers to safeguarding research data.

The idea is 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 – 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.

The point is that this approach is just good preservation practice and provides extra assurance in a way that is:

Simple.

Removes complexity.

Removes dependencies, including lock-in to keeping everything in an EDRM.

Is more manageable over the long-term.

Lowest the risks

Keeps costs down and predictable

WHAT WE DO AT ARKIVUM



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, which means Arkivum has been externally audited for the integrity, confidentiality and availability of data assets in our possession. We follow the OASIS model. We have done our own DRAMBORA risk assessment and we apply ISO16363 principles. We haven't been audited as a Trusted Digital Repository just yet, but that's because accreditation for auditors is still being set up. What we have done is all the hard work on following digital preservation standards and this takes some 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. This is part of a wider escrow model that includes the software and processes we use to run the service as well as ring-fencing of money for fully paid up long-term retention contracts.

100% data integrity guarantee

- All data is returned 'bit perfect'
- No restriction on time
- No restriction on volume
- Included in the SLA
- Worldwide insurance backed: £5M per loss event

© Arkivum Ltd 2013

ARKIVUM

Good preservation practice based on data tape technology, trained staff and very carefully controlled processes also means we can offer a guarantee of data integrity.

All data returned from our service is always bit-for-bit identical to the data the customer supplied, with no restrictions on time or volume.

The guarantee is backed by insurance, is included in our SLA.

You might want to compare this with other cloud storage providers

Data escrow

- Copy of data offline at third-party escrow site
- LTFS on LTO tape with open source tools
- Customer has access to escrow copy:
 - If we fail to provide the service
 - If the customer decides to leave



© Arkivum Ltd 2013


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.

Use of LTO and LTFS with 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.

Pricing

- PAYG or Paid-Up for 5,10 or 25 year
- No ingress or egress charges
- Migrations and refreshes included
- Audits and certification included
- Escrow copy included, no exit costs



© Arkivum Ltd 2013 ARKIVUM

The pricing model for the service is simple.

We support both opex, i.e. PAYG, and capex, i.e Paid-Up, models. So, for example, we can offer retention for 10 years for a fixed up front cost. And, unlike other storage service providers, there's no charges for getting data in and out of our service.

And, the price includes all those migrations and refreshes, and it includes us undergoing 6 monthly ISO27001 audits and our own internal assessments, and it includes the escrow copy of the data, which the customer can take away with them should they decide to leave the service.

So the important thing is that the cost of long-term retention is predictable and cost-effective, it can be fixed, and it includes all the actions needed at the file level for proper preservation.

Thank you

- Arkivum Stand
- www.arkivum.com
- matthew.addis@arkivum.com

"Arkivum has helped us to create a robust archiving solution that will allow us to focus our budget on the business rather than yet more storage"

"Archived documents can then be seamlessly accessed from within the document management system, in the same way current documents are".