ARKIVUM
ASSURED ARCHIVING

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

Challenges and solutions to long-term records retention and access

Dr Matthew Addis
Arkivum Ltd
SAG Edinburgh 2013

Contents

- Retention of electronic documents and records
- What are the challenges?
- How can they be overcome?

"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 Unit

WHY KEEP ELECTRONIC DOCUMENTS AND RECORDS?

Why keep electronic documents and records?

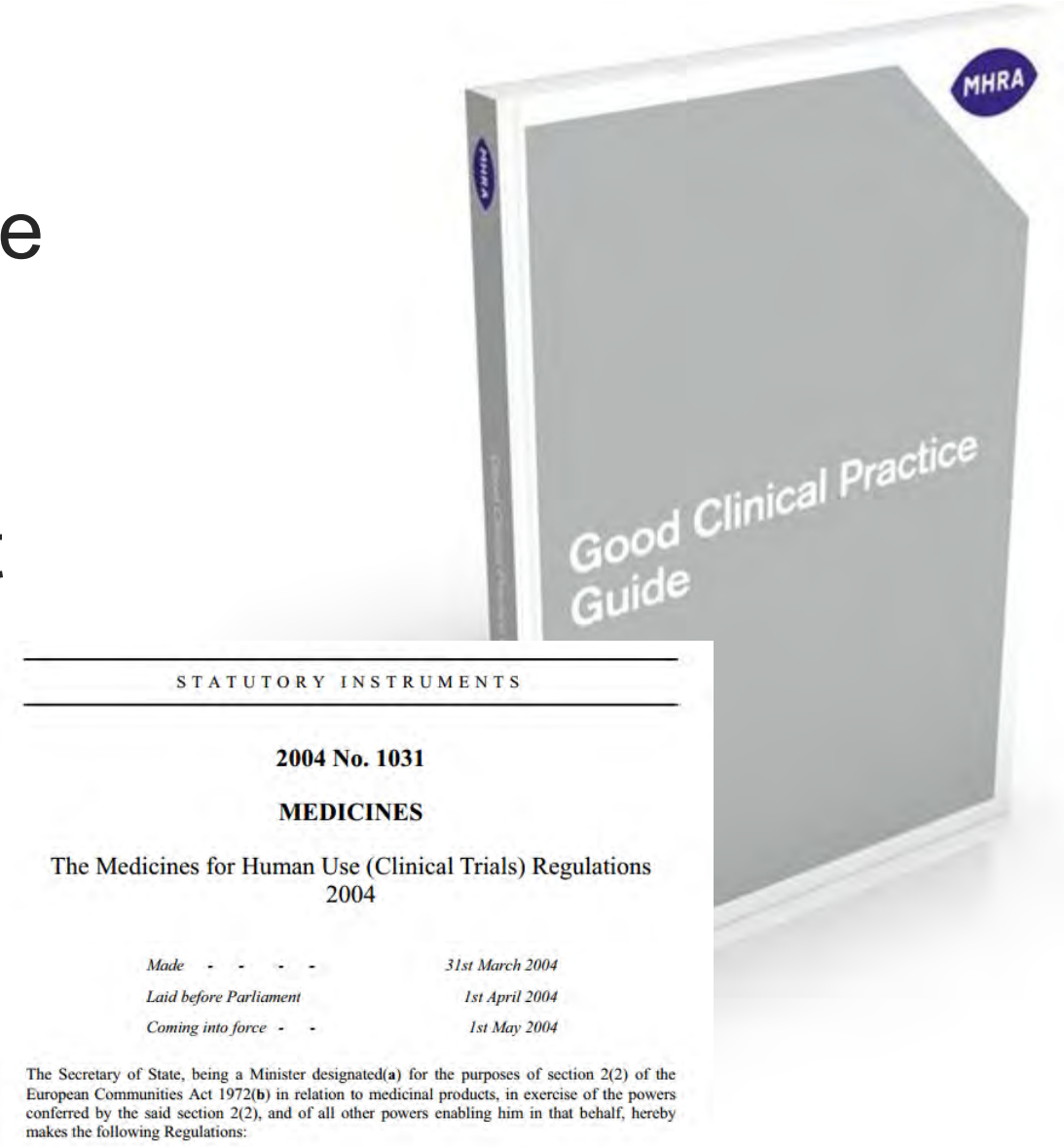
- Regulation and compliance
- Intellectual Property
- Reuse
- Save money
- Better use of in-house resources

GxP Regulations

- FDA 21 CFR Part 11
- EudraLex Volume 4 Annex 11
- OECD GLP

MHRA

- MHRA GCP guide
- Medicines for Human Use Clinical Trials Act (SI 2004/1031)



Research

- MRC GRP guide

Records Management

ISO 15489-1:2001

Information and documentation -- Records management -- Part 1: General

ISO 23081-1:2006

Information and documentation -- Records management processes -- Metadata for records -- Part 1: Principles

MoReq2010®

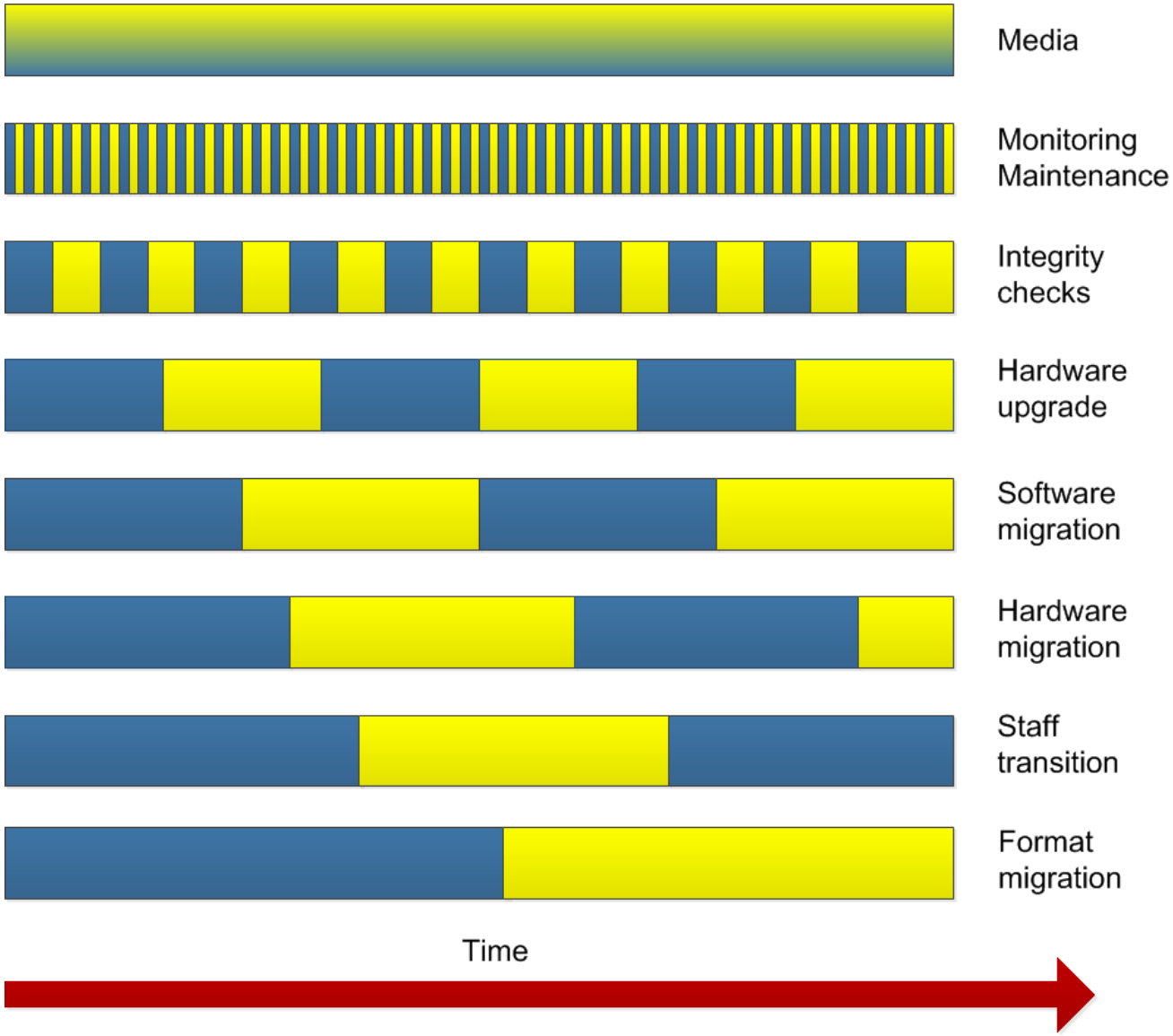
Modular Requirements for Records Systems

Common Requirements

- Integrity
- Authenticity / reliability
- Access control / confidentiality
- Usability / ready access / readability
- Responsibility / named individuals
- SoPs, validation, audit trails
- ***Risk management***

WHAT ARE THE CHALLENGES?

20 years of keeping content alive



Digital Preservation

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

Jeff Rothenberg

Perpetual Change

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

Change and Risk

- Do nothing
- Do the wrong thing
- Delayed action
- In-house
- Use a service provider

Risks

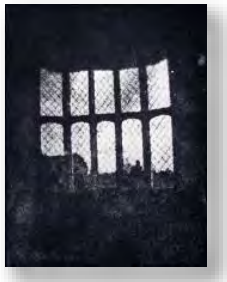
- Technical obsolescence, e.g. formats and apps
- Hardware failures, e.g. digital storage systems
- Loss of staff, e.g. skilled archive and IT staff
- Insufficient budget, e.g. storage too expensive
- Accidental loss, e.g. human error
- Selection, e.g. don't retain what you should
- Stakeholders, e.g. retention no longer a priority
- Underestimation of resources or effort
- Fire, flood, meteors, aliens…

Risk Identifier:	R55	
Risk Name:	Loss of integrity of information	
Risk Description:	Repository is incapable of demonstrating that the integrity of information has been maintained since its receipt, and that what is stored corresponds exactly with what was originally received.	
Is this Risk Relevant?:	• Does repository commit to preservation of information integrity?	
Example Risk Manifestation(s):	• Records documenting government expenditure have been subjected to unauthorised or unanticipated changes, rendering them no longer representative of originally deposited content	
Nature of Risk:	Physical environment	
	Personnel, management and administration procedures	
	Operations and service delivery	X
	Hardware, software or communications equipment and facilities	
Owner:	Preservation	
Escalation Owner:	Preservation	
Stakeholders:	Management; financiers; staff; depositors; users; producers	
Mitigation strategy(ies):	Avoidance strategies: • Ensure policies and procedures are conceived with due consideration of integrity requirements • Maintain and review policies and procedures to ensure adequate recording and comparison of checksums to demonstrate that archived information has suffered no loss of integrity since its deposit or receipt • Ensure software and hardware systems and preservation strategies are capable of preserving information integrity In the event of risk's execution: • Invoke treatment strategies to alleviate loss of reputation or trust	
Risk Relationships:	→R01 [contagious] →R02 [contagious]	
Risk Probability:	4	
Risk Potential Impact:	3	
Risk Severity:	12	

30,000 years

2,200 years

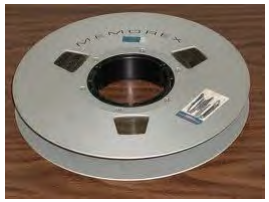
1,200 years



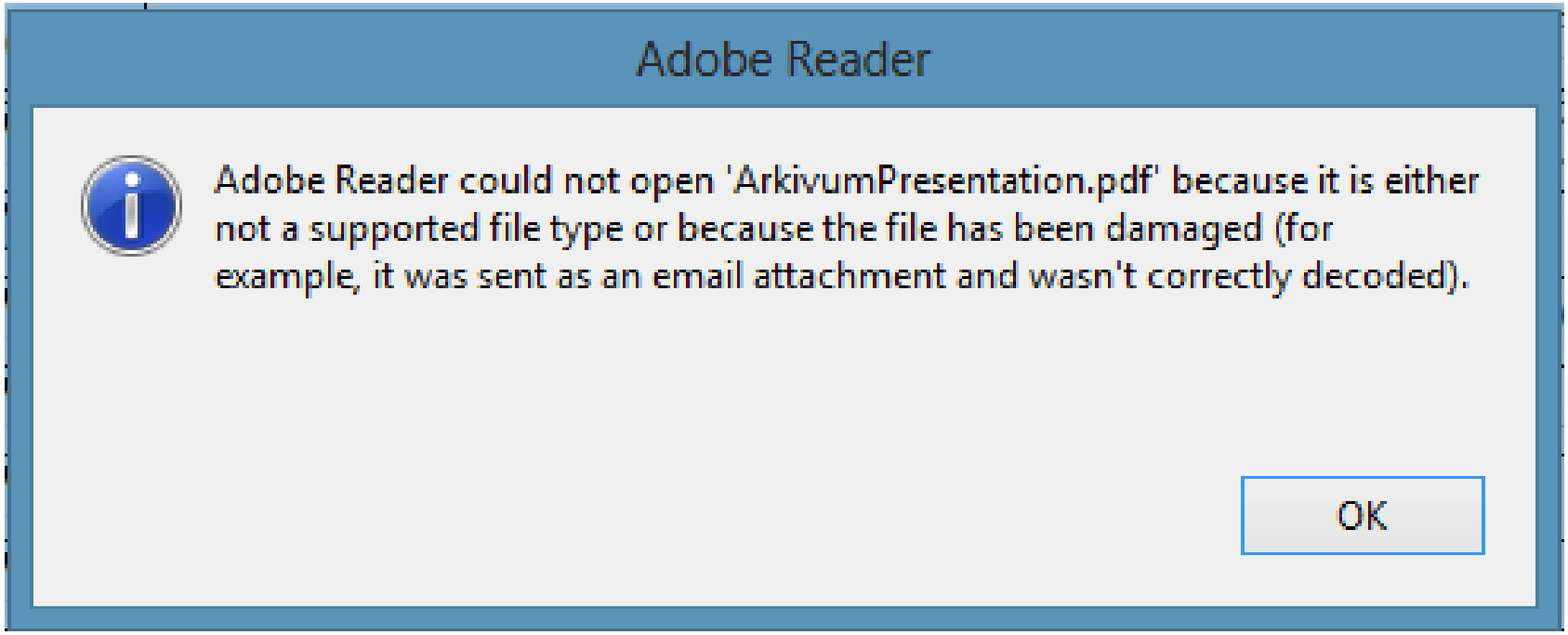
200 years

Each new generation of technology:

1000x times denser
lasts 1/10th as long



20 years



It's not getting any better!

- 1000 times more HDD capacity over last 15 years
- Only 10 times lower Bit Error Rates (BER)

- HDD BER = 10^{-14}
- 1 TB = 10^{13} bits
- 10% chance of an error when reading all of a HDD

- *Within a few years, you are more likely than not to get a read error when copying a HDD*

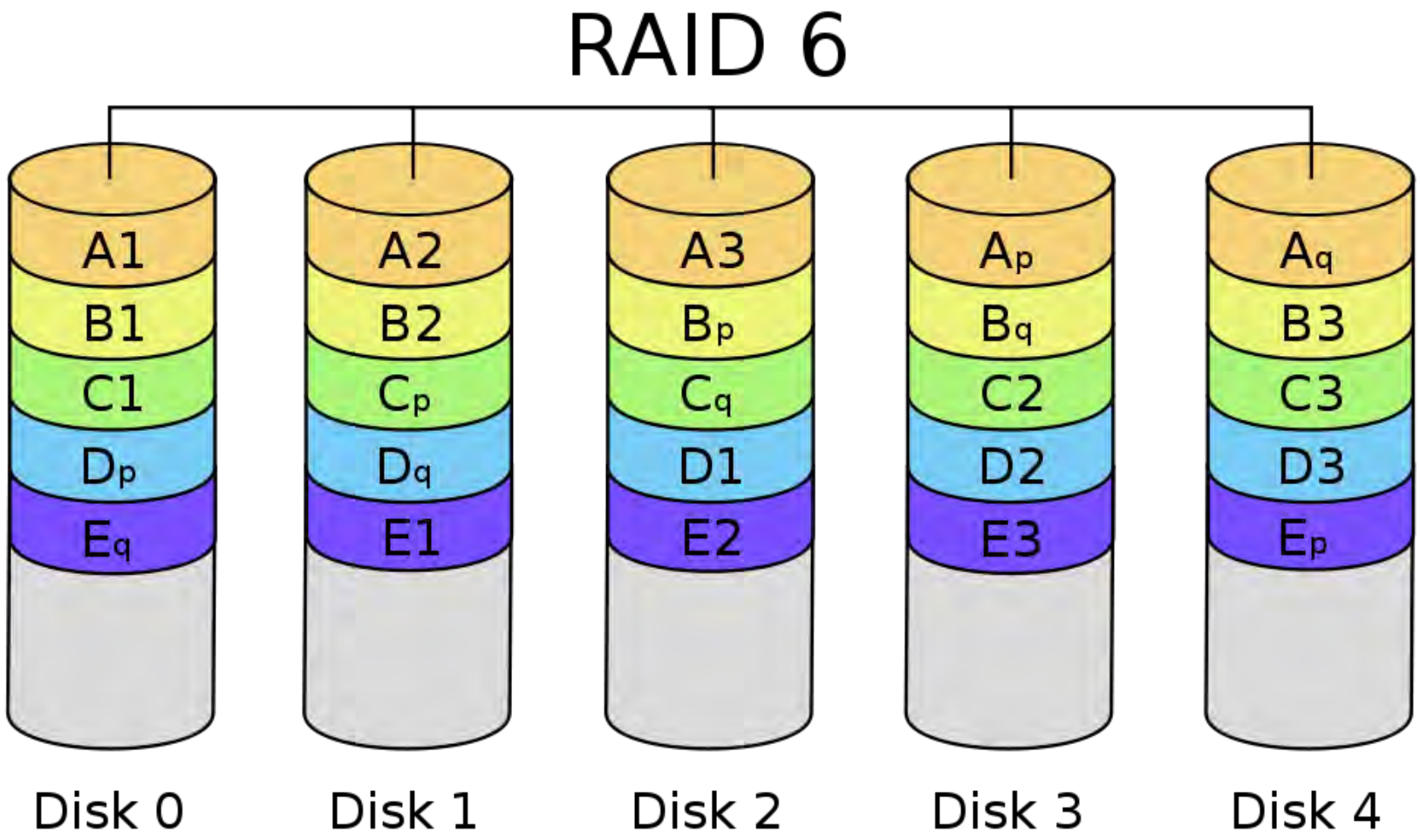
No storage media is 100% safe

- SSD
- CD, DVD
- Data Tape

Examples of damaged discs.

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

The IT Industry knows this already

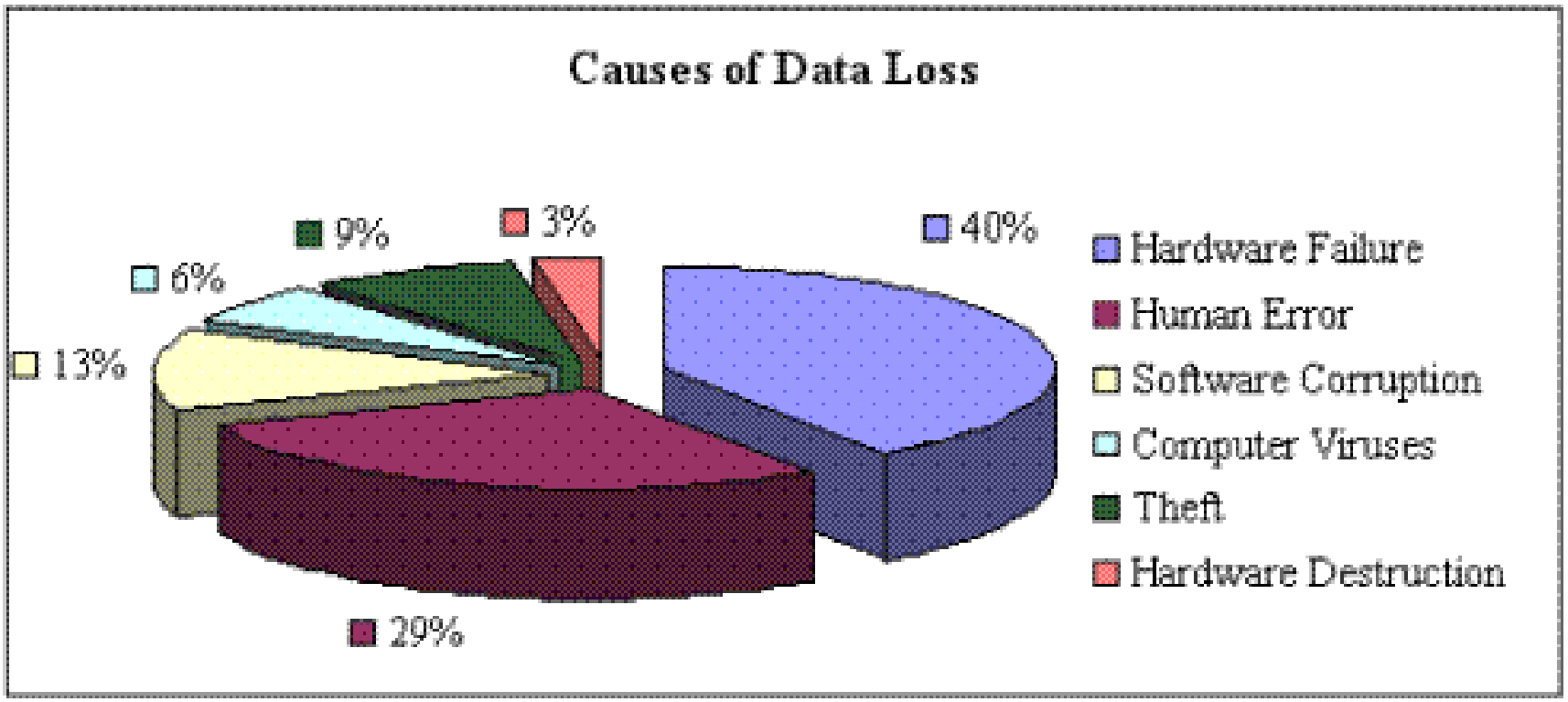


Systems bring their own problems

"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"

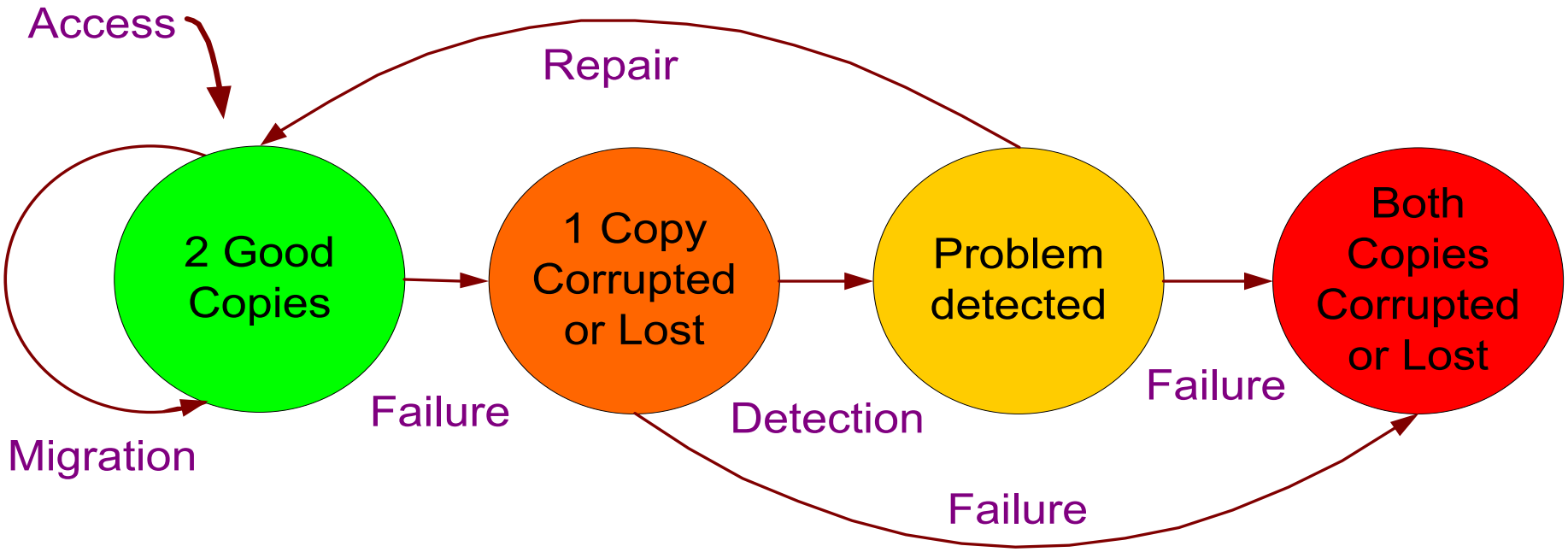
NetApp study of 1.8M HDD in 155,000 systems

People cause data loss too!



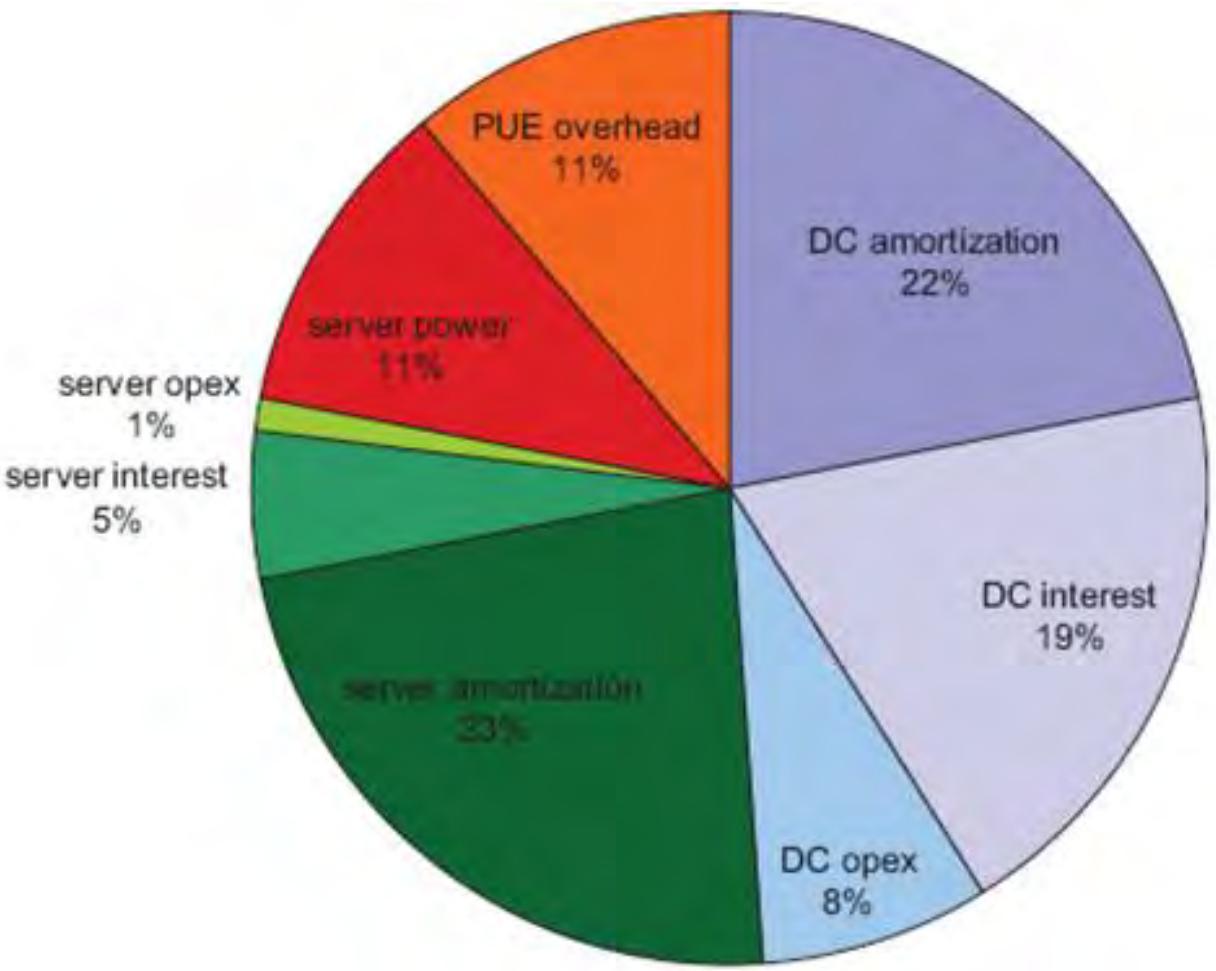
Drive risk down

- Diversity keeps things safe
- Active management of data integrity
- Migration to address obsolescence



Why not make more copies?

Storage TCO



Google

Data

- Documents: eTMF, eCTD
- Supporting data
  - LIMS
  - NGS
  - MRI, CT, 3D ultrasound, 2D x-ray
  - Sensors and monitoring
  - Large cohort studies

- Large *numbers* of files and/or large volumes

Something has to give

- >100% CAGR in data volumes / records
- 20% CAGR in HDD capacity per £ IHS iSuppli
- 2% CAGR in IT Budgets ComputerEconomics

→ Store less

→ Compress

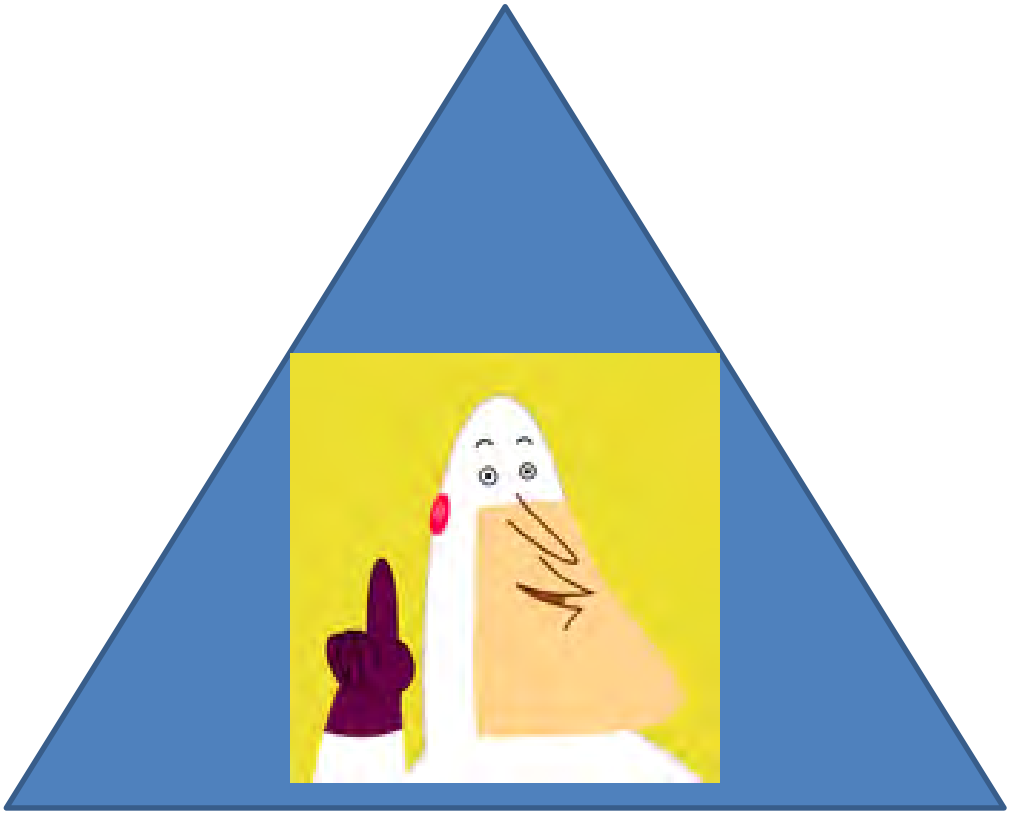
→ Archive outside of EDM systems

MAKING CHOICES

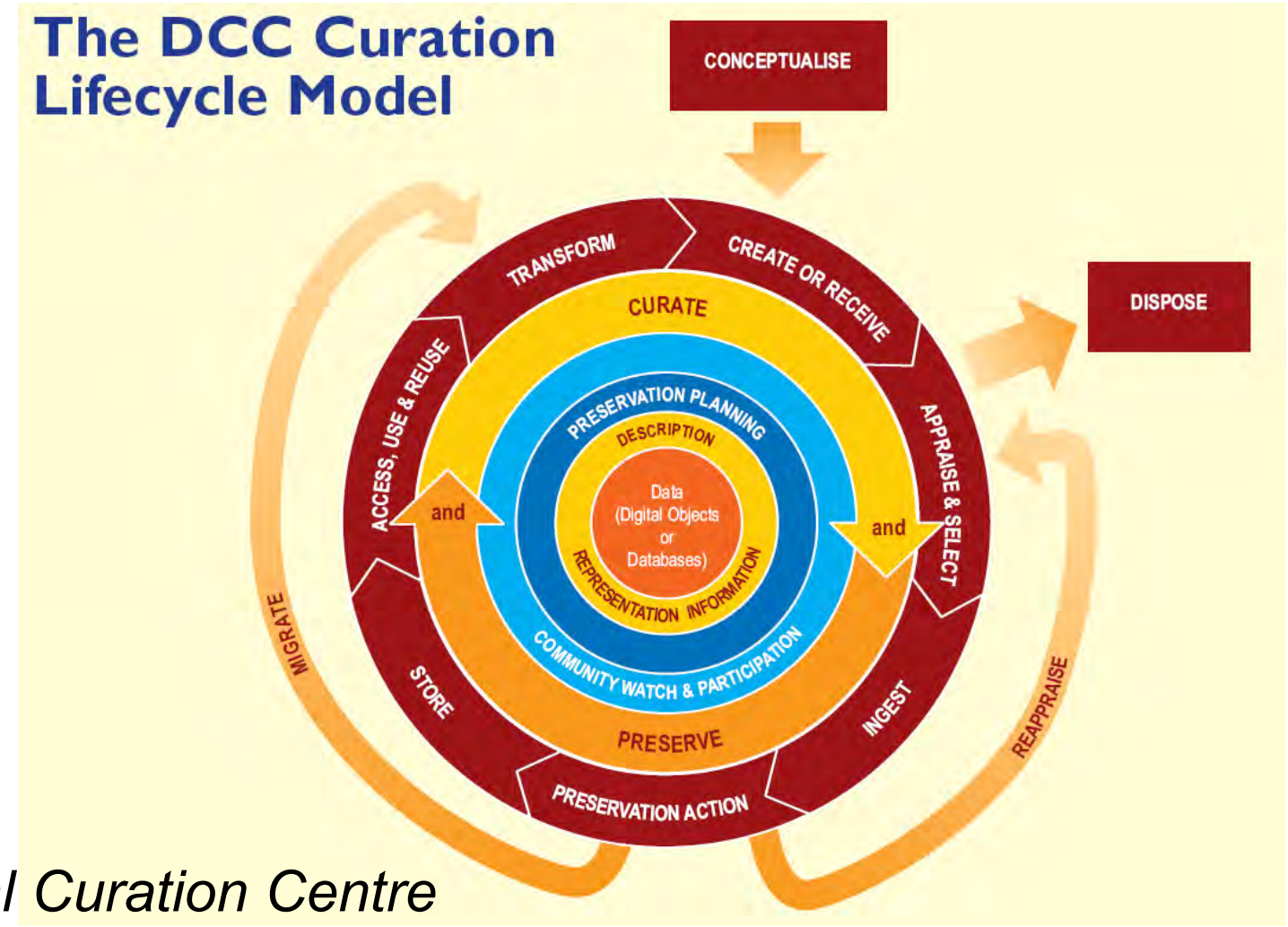
Cost v.s. safety v.s. access

	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)

How do you make a choice?

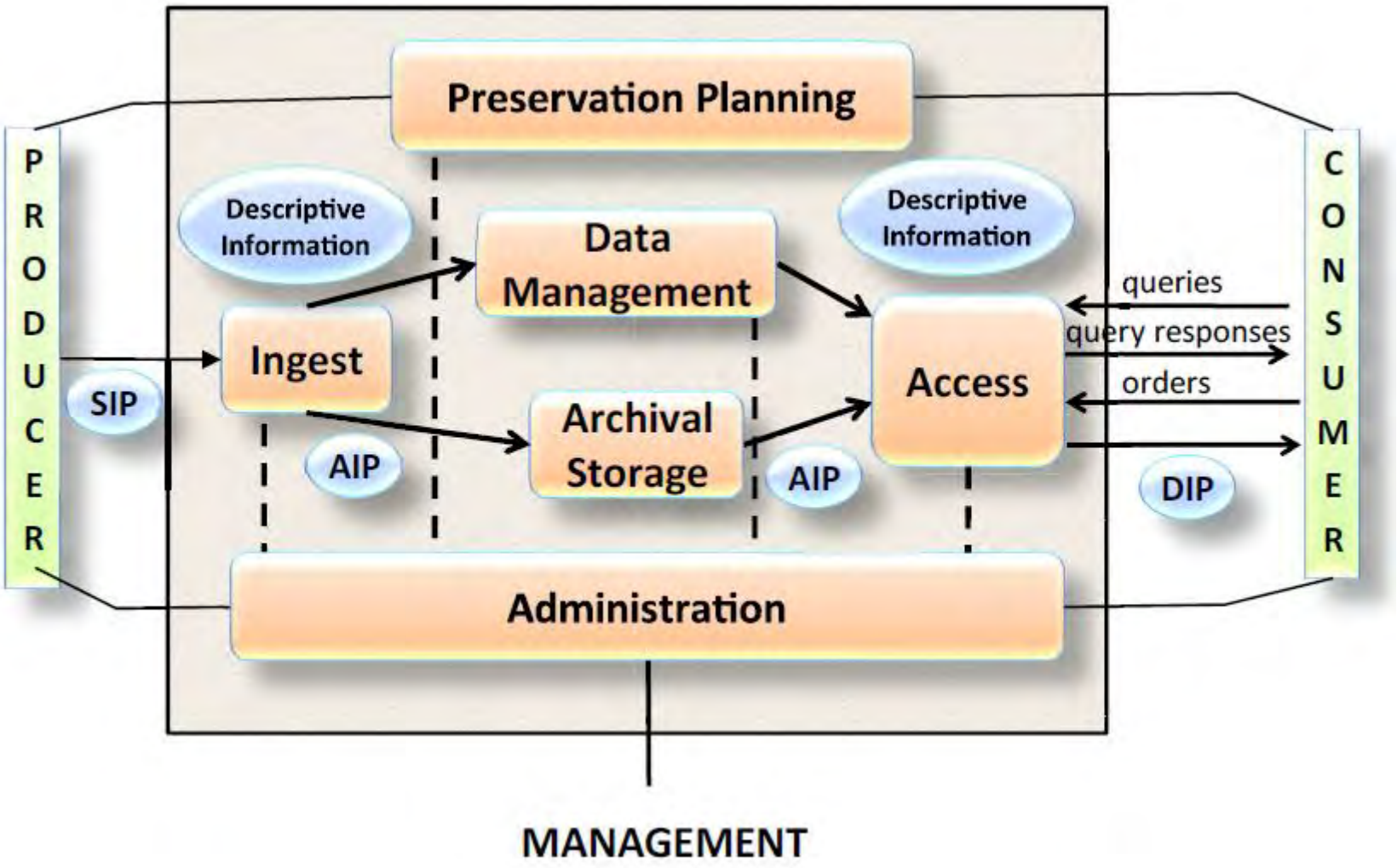


Digital Content Lifecycle



Digital Curation Centre

ISO 14721 (OAIS) 2012



Metadata standards

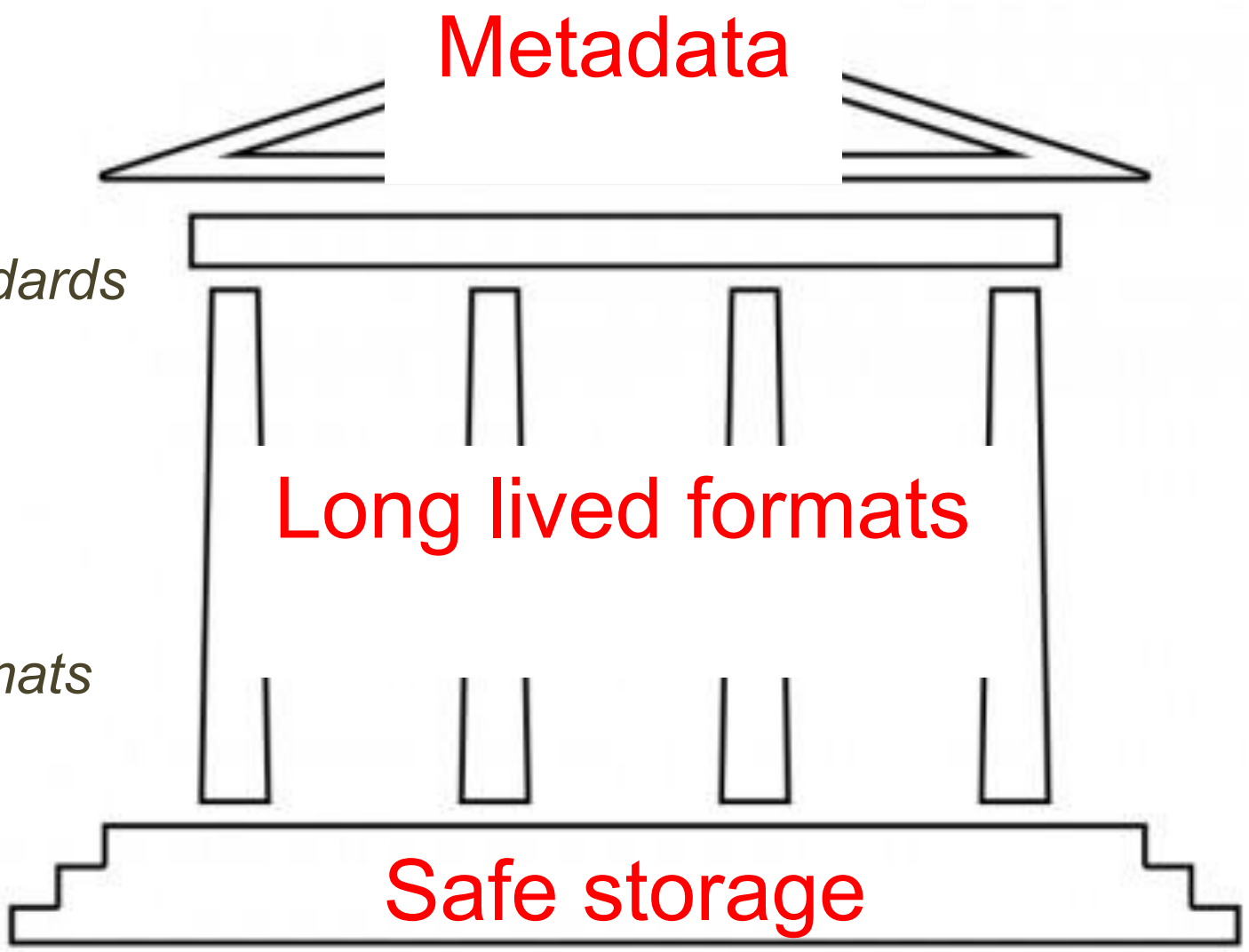
ISO 16363 (TDR) 2012

DRAMBORA interactive — Digital Repository Audit Method Based on Risk Assessment

A SIMPLE STRATEGY

Simple strategy

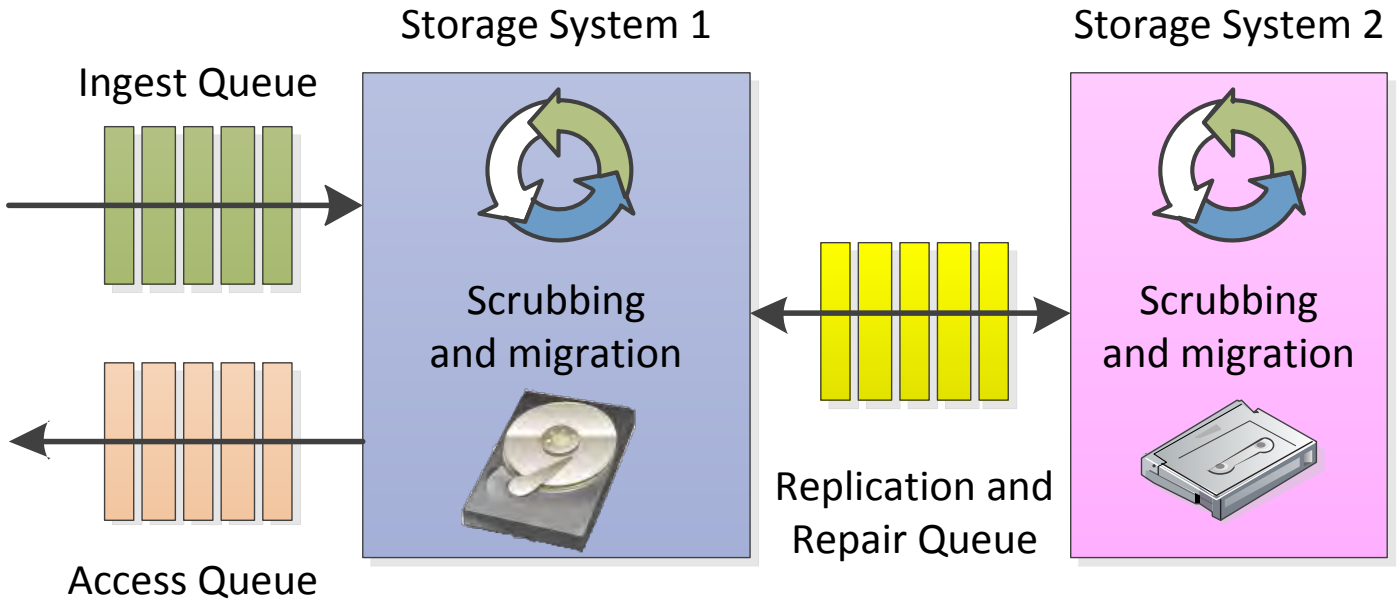
Descriptive
Technical
Preservation
Rights
Metadata Standards



Creation
Preservation
Access
Canonical Formats

Optimise for archive

Active archiving



- Preservation best practice (diversity, intervention)
  - Multiple copies in different locations
  - Different technologies and different people
  - Active management: migration, integrity

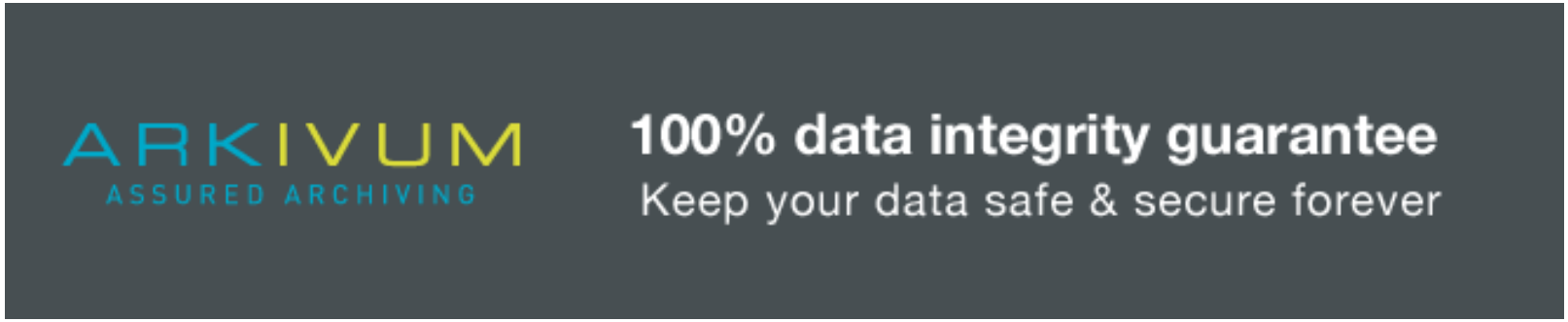
Burnline

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

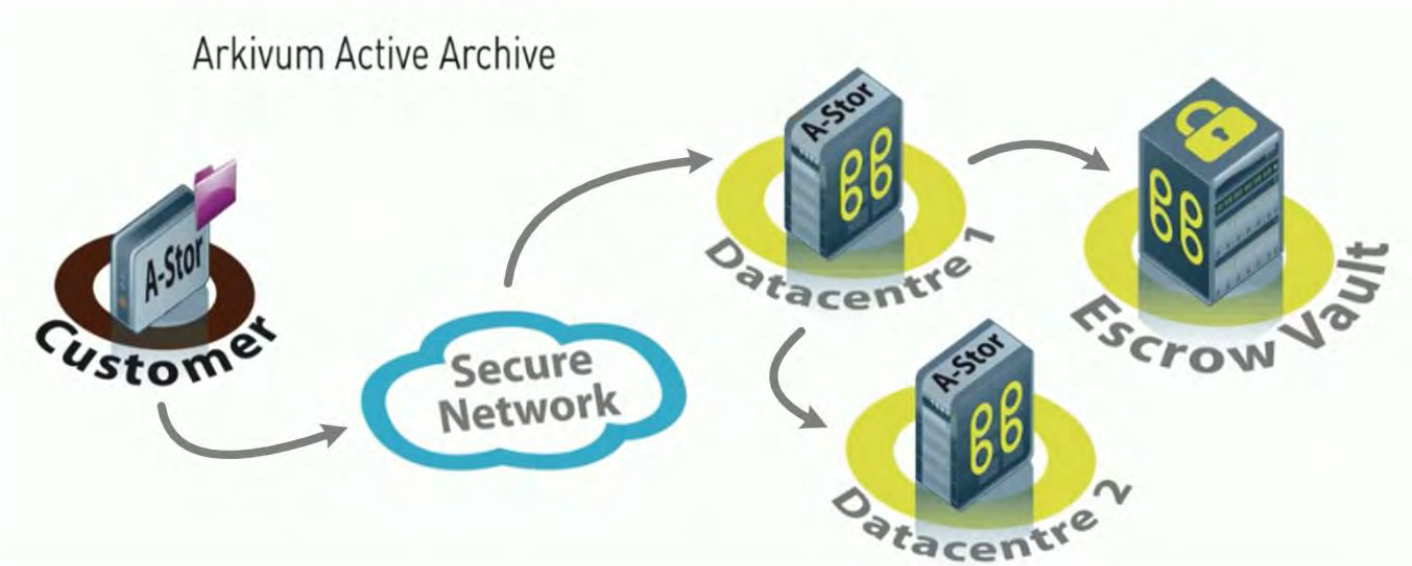
HOW ARKIVUM DOES IT

Arkivum

- Online data archiving as a service
- Spin-out of University of Southampton
- Decade of know-how working with archives
- Safe, secure, accessible data storage
- Designed from ground-up for retention and access



Approach



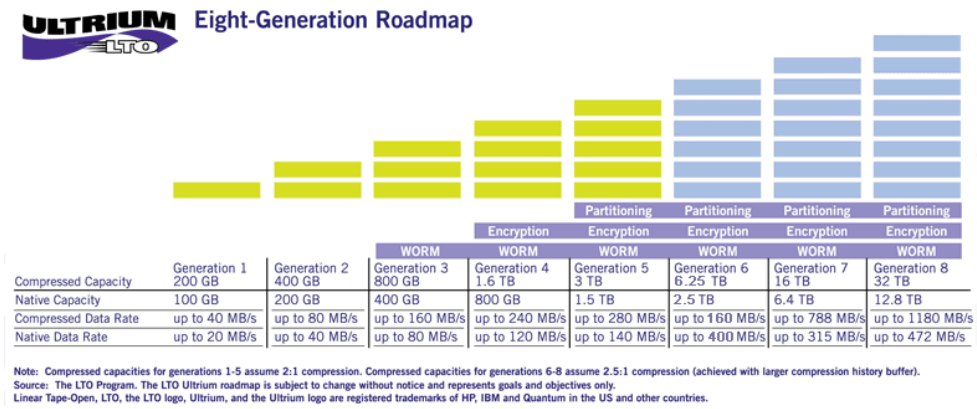
- Best practice: 3 copies, 3 locations, online and offline
- Active archiving: integrity, obsolescence
- Risk management: ISO27001, Drambora, TDR

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

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



	Generation 1	Generation 2	Generation 3	Generation 4	Generation 5	Generation 6	Generation 7	Generation 8
Compressed Capacity	200 GB	400 GB	800 GB	1.6 TB	3 TB	6.25 TB	16 TB	32 TB
Native Capacity	100 GB	200 GB	400 GB	800 GB	1.5 TB	2.5 TB	6.4 TB	12.8 TB
Compressed Data Rate	up to 40 MB/s	up to 80 MB/s	up to 160 MB/s	up to 240 MB/s	up to 280 MB/s	up to 160 MB/s	up to 788 MB/s	up to 1180 MB/s
Native Data Rate	up to 20 MB/s	up to 40 MB/s	up to 80 MB/s	up to 120 MB/s	up to 140 MB/s	up to 400 MB/s	up to 315 MB/s	up to 472 MB/s

Note: Compressed capacities for generations 1-5 assume 2:1 compression. Compressed capacities for generations 6-8 assume 2.5:1 compression (achieved with larger compression history buffer).
Source: The LTO Program. The LTO Ultrium roadmap is subject to change without notice and represents goals and objectives only.
Linear Tape-Open, LTO, the LTO logo, Ultrium, and the Ultrium logo are registered trademarks of HP, IBM and Quantum in the US and other countries.

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

Thank you

- 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".