Preservation at scale

Arkivum
Every bit archived
Arkivum in 60 seconds

SLA with 100% data integrity guaranteed

World-wide professional indemnity insurance

Long term contracts for enterprise data archiving

Fully automated and managed solution

Audited and certified to ISO27001

Data escrow, exit plan, no lock-in

© Arkivum 2014
Example Customers

- FCA
- TATE
- OXFORD FERTILITY UNIT
- RGB Building Supplies
- NEWVOICEMEDIA
- Neilson Financial Services
- PHLEX
- OXFORD Molecular Diagnostics Centre
- Queen Mary University of London
- Queen's University Belfast
- The University of York
- University of East London
- UNIVERSITY OF LEEDS
- UCL
- University of Hertfordshire
- Loughborough University
- UNIVERSITY OF SOUTHAMPTON
- Aston University

© Arkivum 2014
Some ways to look at scalability

• Capacity, Objects, Customers, Users
  ➔ Throughput (ingest)
  ➔ Latency (access)
  ➔ Scale-out (elastic)

• Cost, Duration
  ➔ Economies of scale
  ➔ DCF, Capital Expenditure, Endowments

• Data Integrity and Risk Management
  ➔ Chain of custody
  ➔ Exit plan
Scalability: IT perspective

- Data volume
- Storage unit cost
- IT budget
Scalability: data integrity

- 99.99% chance of making a car journey without breaking down

But…
- 600 trips per person per year in the UK
- 20,000 breakdowns per day

- 10,000,000,000 data objects
- Each object makes 1,000 ‘data trips’
- 20 year total data journey
- 100% guarantee of no permanent breakdowns
Scalability: integrity through time
Active Archiving

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

A-Stor creates an encrypted copy on the appliance and copies to Arkivum’s Data Centre One.

Once all three copies are confirmed safe and secure the original copy can safely be deleted.

A-Stor DC creates the escrow copy and stores it with a third-party.

A-Stor DC creates the second copy in Arkivum’s Data Centre Two.

Secured Network
But…

• How do you know everything is correct?
  – Missing data
  – Extra data
  – Corrupted data
  – Metadata about the data (technical, preservation)

• Are all the processes validated?
  – End-to-end chain of custody

• What tests and verification can be done?
  – Trust but verify!
Archive File Packages (1)

Original data

- Archivematica
- ePrints
- DSpace
- Figshare
- Manual process

AIPs
- Folders of files
- Single files

Receive Files

Validate Files

Encrypt, chunk, aggregate

Validate Archive File Package

Checksums, file metadata

XML file descriptions

Chunk data

Bagit structure

Original data
Archive Packages (2)

1. Upload to DCs
2. Validate AFP
3. Write to tape
4. Validate AFP
5. Send to Escrow
6. Confirm Escrow
7. Export list of ‘safe data’
8. Notify appliance that AFP is safe
9. Delete original data
End-to-end

• Starts and ends with the customer
  – Chain of custody for data ‘on the way in’
  – Never delete anything unless confirmed in escrow
  – All data checked regularly and ‘on the way out’

• Customer can validate escrow tapes

LTFS

OpenSSL

W3C

© Arkivum 2014
Back to scalability

- Archive Packages help us with scaling up
  - Self describing metadata and data
  - Multiple files and folders
  - Can detect additions, omissions, corruptions
  - Easier to handle than individual files
  - Based on open specifications, e.g. Bagit
  - Easier to replicate, migrate, track, check
  - Support a scalable customer exit strategy
But not the end of the story

CCSDS
The Consultative Committee for Space Data Systems

Recommendation for Space Data System Practices

AUDIT AND CERTIFICATION OF TRUSTWORTHY DIGITAL REPOSITORIES

A risk-aware path to self-assurance and partner confidence for digital repositories

http://www.repositoryaudit.eu/

© Arkivum 2014
Preservation at scale is more than technology:

- Skilled and trained people
- Validated processes and procedures
- Comprehensive risk management
- Audit and validation
- Specialist infrastructure
- Economies of scale and predictable costs
- End to end chain of custody
- Exit strategy