RISING TO THE IMPOSSIBLE CHALLENGE

THE STORAGE OF ALL RESEARCH DATA....

....FOREVER??

Tim Banks, University of Leeds
Wendy White, University of Southampton
Matthew Addis, Arkivum

Photo Credit: PNNL - Pacific Northwest National Laboratory via Complight cc
ARCHIVE IN THE CLOUD?

- Efficient at scale
- Flexible
- POSF* cost recovery
- Little local effort required
- No large capital investment needed
- Data escrow
- Finance escrow (for POSF)
- Purchase Tax (e.g. VAT)
- Bandwidth
- Data security
- Ingress/egress charges
- Storage location

*Pay once, store forever (25yrs)

Photo Credit: kevin dooley via Compfight cc
<table>
<thead>
<tr>
<th>Item</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base frame inc. drives, robot etc.</td>
<td>€146,950</td>
</tr>
<tr>
<td>5yrs support for base frame</td>
<td>€55,250</td>
</tr>
<tr>
<td>Professional Services (installation)</td>
<td>€8,250</td>
</tr>
<tr>
<td>LTO-6 media inc. cleaning tapes</td>
<td>€32,750</td>
</tr>
<tr>
<td>120TB Spinning Disc cache</td>
<td>€50,000</td>
</tr>
<tr>
<td>5yrs support for disc cache</td>
<td>€22,250</td>
</tr>
<tr>
<td>QStar Archive manager software</td>
<td>€215,100</td>
</tr>
<tr>
<td>Infrastructure (fibre channel switches)</td>
<td>€31,500</td>
</tr>
<tr>
<td>1x full time archive manager post (5yrs)</td>
<td>€308,000</td>
</tr>
<tr>
<td>Machine Room costs; power, cooling (5yrs)</td>
<td>€123,200</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>€993,250</strong></td>
</tr>
</tbody>
</table>

(€176.50 per TB per year for single copy)

1.125PB raw; 450x LTO6; 4x drives; 1x robot; archive management software; 120TB disc cache

Photo Credit: naezmi via Compfight cc
The Data Archive Cloud Model

Data Archive as a Service

Submission of data for ingest

Request data for retrieval

Photo Credit: infocux Technologies via Compfight cc
For same annualised cost per TB per year, with this storage profile, cloud service is **26% less expensive**

Assumptions:
- Archive manager is employed full time for 5-years
- Any fixed maintenance costs are paid up-front for 5-years
Integration with Repository

1. User requests dataset
2. Repository interrogates archive appliance
3. Dataset held in local cache?
Integration with Repository (2)

Dataset held in local cache?

Yes

Provide download link

No

Present user with request form
Integration with Repository (3)
Integration with Repository (4)

1. Repository sends request for retrieval of dataset
2. Dataset retrieved and verified
3. - Request form replaced with download link
   - Email link to user
<table>
<thead>
<tr>
<th>ePrints</th>
<th>Arkivum</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data Deposit</strong></td>
<td><strong>Data Request</strong></td>
</tr>
<tr>
<td>Data file + metadata deposit by researcher</td>
<td>Metadata view &amp; data file request by reader</td>
</tr>
<tr>
<td>URI generation</td>
<td>Data file checksum + encryption and decryption</td>
</tr>
<tr>
<td>Data review by ePrints Editor</td>
<td>Data file request move to/from Arkivum</td>
</tr>
</tbody>
</table>
| DOI request:  
  - Data meets DOI policy requirements  
  - DOI minted | Data deletion available when secure storage transfer completed (can be automated) |
| Faculty Data review RDM policy 6.3.1 | Checksum validation of data by Editor (can be automated) |

**Data Storage**
- ISO 27001 certified
- > 10 year storage
- Automated checks for data integrity

**Secure storage**
- at Data Centre 1
- at Data Centre 2
- on tape at escrow facility

Julian H. Ball; Dorothy R. Byatt (2014)
Final Considerations

• It is important to fully understand the real costs of running a local archiving service
• Consider where your data is stored (Data Protection)
• Understand the full costs of cloud archiving service (including ingress and egress costs and limitations).
• Look at your likely 5-year archive volume profile
• Understand the protection level that you require
• Have an exit strategy