Digital Preservation
In Practice

Ed Fay
London School of Economics

EMAIL: e.fay@lse.ac.uk
TWITTER: @digitalfay
We do not have a long-term digital preservation strategy
We do not have a long-term digital preservation strategy

...and we’re perfectly ok with that!
This does mean....

That we are agnostic about the ‘final’ solution

(We don’t even necessarily think there is/has to be one...)

That we are happy to investigate, experiment and change what we are doing
This doesn’t mean…

We are doing nothing!
We have:

Decided we want to be preserving digital stuff

We are:

Starting to take the first steps
Long-term means long-term

- Digital preservation is about preserving access at a given point in time
- Right now we have no critical collections but, given time, we will
Long-term means *long*-term

- We don’t need:
  - a ‘complete solution’; right now
- We do need:
  - to start thinking, getting clear about the problem, and talking about it
- We also need to start doing something so it doesn’t become too late
What have we done?

• Collections audit (a spreadsheet)
• Risk assessment (DRAMBORA)
• ...
• User requirements analysis (ongoing, for curators as well as end-users)
<table>
<thead>
<tr>
<th>Collection</th>
<th>Identifier</th>
<th>Format</th>
<th>Size (MB)</th>
<th>Media</th>
<th>Location</th>
<th>Accretion/Attrition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exam papers</td>
<td></td>
<td>PDF</td>
<td>1126</td>
<td>Server/Tape Backup</td>
<td><a href="https://library-2.lse.ac.uk/protected-exam/">https://library-2.lse.ac.uk/protected-exam/</a></td>
<td></td>
</tr>
<tr>
<td>Research publications</td>
<td>LSERO</td>
<td>PDF</td>
<td>4710</td>
<td>Server/Tape Backup</td>
<td><a href="http://eprints.lse.ac.uk">http://eprints.lse.ac.uk</a></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Word/data</td>
<td></td>
<td>Server/Tape Backup</td>
<td><a href="http://eprints.lse.ac.uk">http://eprints.lse.ac.uk</a></td>
<td></td>
</tr>
<tr>
<td>e-Theses</td>
<td>LSETO?</td>
<td></td>
<td></td>
<td>Server</td>
<td>lib-6</td>
<td>stable (once it has been migrated)</td>
</tr>
<tr>
<td>Data Resource Archive</td>
<td></td>
<td>text/PDF/SBR</td>
<td>~5000?</td>
<td>Server</td>
<td>lib-6</td>
<td></td>
</tr>
<tr>
<td>Paddy Ashdown diaries</td>
<td>M3348</td>
<td>Word/html/en</td>
<td>1628</td>
<td>CD (4)</td>
<td>Archives</td>
<td>Expecting more Ashdown</td>
</tr>
<tr>
<td><strong>Risk Number 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Risk Name:</strong></td>
<td>Loss of trust or reputation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Risk Description:</strong></td>
<td>One or more stakeholder communities have doubts about the repository's ability to achieve its objectives.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Risk Areas:</strong></td>
<td>Personnel, Management &amp; Admin Procedures, Operations &amp; Service Delivery</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Vulnerability(ies):</strong></td>
<td>* An irrecoverable loss of digital objects provokes community concerns about the repository's competence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Consequence(s):</strong></td>
<td>* Credibility of the organisation as a location for deposit of digital objects, and associated funding, reduced.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Relationship(s):</strong></td>
<td>No Relationships Established</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Risk Owner(s):</strong></td>
<td>Senior Manager</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Tools: DRAMBORA

- Risk assessment
- Why?
- Start the conversation
- Make the problems clear to all stakeholders (curators, technical specialists, senior managers)
- Not for detailed functional analysis
- http://repositoryaudit.eu/
What are we doing?

• Creating a place to store our digital objects
  • ‘repository core’ for object storage (redundancy, backups) and identification

• Creating a way to ingest/accession objects
  • workflow for object characterisation, checksums, quarantine/virus check
Tools: Archivematica

- Workflow tool
- Why?
- For ‘ingesting’ digital objects
- Bundles tools for:
  - quarantine/virus check (filesystem, ClamAV)
  - checksum creation/verification (MD5)
  - format characterisation/validation (FITS which packages DROID, JHOVE, NZ Metadata Extractor ...)
- http://archivematica.org/
What will we be doing?

- Building management interfaces
- Building access interfaces
- Developing a logical preservation approach
  - strategy and policies
- Building logical preservation functionality
  - implementing tools for migration or emulation
LSE Digital Library

Curation

Ingest
(Create objects)

Management
(Update, delete objects)

Repository Core
(Digital object storage, identification)

Access

Web applications
(Presentation, rendering)

Search
(Indexing, retrieval)

Access Control

Bitstream preservation
(Integrity checking, backups)

Logical preservation
(Format migration, emulation)
LSE Digital Library: Design Principles

- Flexible—we can hold a range of different types of digital collection
- Extensible—we can adapt to changing collections and user requirements
- Modular—we can replace components without disrupting other functions
Tools

- **DRAMBORA**
  - *used* for high-level risk assessment – as a tool for starting the conversation

- **Archivematica**
  - *used* to characterise our collections and to assist in producing a more detailed risk profile for further analysis

- **Fedora/Hydra repository**
  - *will be used* to store/manage all our digital collections

- **Planets/Plato**
  - *will(?) be used* to help us plan our long-term strategy
Guiding principles

• Openness: standards, technologies
• Transparency: clear, documented decisions and processes
• Engagement: bringing everyone along with us (senior managers, depositors, colleagues across the library)
The **problems**…
...you’ve heard about

The **solutions**…
...are complex

*Digital Preservation is HARD!*
Digital Preservation is HARD!
OAIS can help

But it can also scare people!
Digital Preservation is HARD!

Shorter summary of DP: know what you have and value, assess risk, take action to avoid risk, repeat.

Problem: people don't do it

Steve Hitchcock
JISC KeepIt Project Manager
http://twitter.com/#!/jisckeepit/status/25530206525591552
Small steps now...

- ...save big(ger) problems later
- **Learn by doing** – you (should) always have originals
- Focus on:
  - “*ingest*” = capture, identify
  - “*bit-preservation*” = redundancy, backups
- Use this as the basis for more **thorough risk assessment**
- Use that as basis to **make the case for investment**
- Then think about the long-term (policy and tech)
That’s it!

Ed Fay
London School of Economics

EMAIL: e.fay@lse.ac.uk
TWITTER: @digitalfay