FLASHBACK!

Rescuing Disc-Based Content from the 1980’s to the current day

Maureen Pennock  Peter May
Head of Digital Preservation  Technical Architect

IDCC Wednesday 24th February 2016
Flashback Proof of Concept Project Scope

• Devise bit-level preservation process for legacy handheld material stored on disc:
  – CD
  – DVD
  – 3.5” floppy disk
  – 5.25” floppy disk

• Test migration and emulation workflows to deliver authentic representations of content into reading rooms

• Make recommendations on turning PoC workflows into BAU processes
Content Extraction aka. Disk Imaging

- Twelve step generic workflow, incl.:
  - Disc clean
  - MD5 checksums
  - Disk imaging
  - Virus checking
  - MD5 verification

- Tested:
  - Bitcurator
  - ISOBuster
  - Kyroflux (5.25” disks only)

- More granular versions of workflow derived for each storage media type
Disc Imaging Outcomes

<table>
<thead>
<tr>
<th>Disc Type</th>
<th>Total Discs</th>
<th>Successes</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDs</td>
<td>22</td>
<td>21</td>
</tr>
<tr>
<td>DVDs</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>3.5&quot; floppy disks</td>
<td>29</td>
<td>27</td>
</tr>
<tr>
<td>5.25&quot; floppy disks</td>
<td>23</td>
<td>20</td>
</tr>
</tbody>
</table>

Legend:
- **Total discs**
- **Successes**
Average Imaging Process Duration

<table>
<thead>
<tr>
<th>Format</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDs</td>
<td>5m50</td>
</tr>
<tr>
<td>DVDs</td>
<td>16m34</td>
</tr>
<tr>
<td>3.5&quot; floppy disks</td>
<td>1m10</td>
</tr>
<tr>
<td>5.25&quot; floppy disks</td>
<td>2m26</td>
</tr>
</tbody>
</table>
Content Preservation

• Compare emulation/migration approach to running on legacy hardware

• Decision tree developed to identify primary preservation pathway to test
  – Used previously defined content categories

• Pathways: Emulate, Migrate, Image-only

• Legacy Lab established with working legacy machines
  – Rendering original items in suitably original environment
Preservation Pathways

• Two main approaches:
  – The bwFLA Emulation-as-a-Service (EaaS) offering from the University of Freiburg
  – Migration workflows, details of which varied depending on the source and target formats.

• Third approach considered:
  – The Interject Solution developed by the British Library as part of the SCAPE project
    • discounted: prototype; limited environment support
Emulation as a Service
The Legacy Hardware Laboratory
## Emulation Results

<table>
<thead>
<tr>
<th>Environment</th>
<th>Emulation tool</th>
<th>#</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apple II</td>
<td>Apple Win (local)</td>
<td>2</td>
<td>Very successful in replicating original behaviour. Quicker keyboard reaction than on original hardware.</td>
</tr>
<tr>
<td>BBC Micro</td>
<td>BeebEm (local)</td>
<td>1</td>
<td>Very successful in replicating appearance/behaviour of s/w. Emulator mimics audio experience by replicating the sound of a disk running</td>
</tr>
<tr>
<td>MS-DOS</td>
<td>DOSBox (local)</td>
<td>6</td>
<td>2 successful</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1 unsatisfactory: content not viewable beyond main menu</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3 incomparable to original: unresolved BASICA dependency</td>
</tr>
<tr>
<td>MS-DOS</td>
<td>QEMU (EaaS)</td>
<td>3</td>
<td>2 successful</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1 unsatisfactory: content not viewable beyond main menu</td>
</tr>
<tr>
<td>Mac 7.0</td>
<td>BasiliskII (EaaS)</td>
<td>6</td>
<td>3 satisfactory</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2 unsatisfactory: Audio problems</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1 unable to run: software dependency issue.</td>
</tr>
<tr>
<td>Mac 9.0</td>
<td>Sheepshaver (EaaS)</td>
<td>2</td>
<td>Both satisfactory</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Note this sample does not contain any sound features</td>
</tr>
<tr>
<td>Windows 3.1</td>
<td>Qemu (EaaS)</td>
<td>8</td>
<td>4 successful with four items</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Issues with 4 items relating to video rendering, graphical glitches/distorted images or audio issues.</td>
</tr>
<tr>
<td>Windows 95/98</td>
<td>Qemu (EaaS)</td>
<td>7</td>
<td>4 successful</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Issues with 3 items relating to video rendering and graphics.</td>
</tr>
</tbody>
</table>
Migration Results

- Small number of migration approaches were tested on text documents and spreadsheets

- Some success in migrating spreadsheets, but formatting loss from Lotus 1-2-3
  - Emulation often yielded better results

- One item could not be opened successfully by any of the migration tools tested (Open Office, LibreOffice and MS Excel)
  - However it rendered using QEMU emulation software

- CSV files opened well in modern Excel (to be expected)

- Video content (e.g. MPEG) successfully rendered on modern Win7 machine
Conclusions

• Demonstrated Emulation and Migration approaches can be used to preserve access to legacy digital collection items
  – Albeit with difficulties

• Assessment of content in the original environment was an excellent way to assess the viability of emulated or migrated versions of the collections.

• 12% of the sample was found to be unreadable at the point of imaging
Conclusions - Decision Tree

• Decision tree useful for directing preservation pathway, however some observations:
  – It is important to base preservation planning decisions on the observed behaviour of content objects
  – It is important to take into account the granularity of content
  – An additional category is needed covering source code and other items where emulation or migration would not be an appropriate approach at present
  – Need to take into account implicit knowledge contained within disk directory structures
Conclusions – Legacy Hardware

• Not always easy to source

• Not always clear of working status of kit when purchasing

• Legacy hardware is temperamental
  – Sometimes it works, sometimes not, and there’s no obvious reason why.

• Rare items can be expensive

• Learning curve for staff unfamiliar with legacy technology
Conclusions - Emulation

- Yielded better results w.r.t. appearance and behaviour than migration
- Older material emulates successfully as long as OS and dependent software available
  - Needs collecting.
  - Freiburg estimate 10-12 environments sufficient for ~95% items
- Emulation doesn’t instantly mean you can use the content. Familiarity with how to use historic operating systems is also required
- Some content requires specific hardware (3D graphics cards) or requires low-latency input for optimal user experience
  - These would benefit from locally installed emulator
Conclusions - Migration

• Migration suitable for some content, particularly where the information content is most important

• (Potentially) not so good where formatting is essential

• Migration dependent on having relevant software capable of understand the original file format
  – E.g. able to migrate from format xyz
Next Steps

• Phase 2
  – Larger sample
    • Better idea of stats (imaging time, failure rates)
    • Better understanding of process
  – Deployment at scale
    • Processes – Emulate? Migrate? Image only?
    • Integration with Library Systems
  – Cost
    – Role of software repository (for software dependencies)
  – Reading Room deployment
  – User Experience
Thank you