Preservation planning at the British Library

Michael Day
British Library
Michael.Day@bl.uk

Presentation outline

• The British Library context
• Approaches to preservation planning
• Collection profiling
• Format assessment
• Summing up
Chapter 1.

Alice was beginning to get very tired of sitting by her sister on the bead and of having nothing to do: once or twice she had peeped into the books her sister was reading, but it had no pictures or conversations in it.
Strategic infrastructure

“Outside of legal deposit, we will prefer to connect to content except in circumstances where the connection is not technically feasible or when we wish to hold and preserve the materials for the long term”

Content Strategy

Digital Preservation Strategy

- Improve our Repository
- Manage risks across lifecycle
- Embed digital sustainability
- Collaborate!
Technical Infrastructure

Storage nodes:
• BL St Pancras (STP)
• BL Boston Spa (BSP)
• National Library of Wales (NLW)
• National Library of Scotland (NLS)

Access gateways:
• Bodleian Library (Ox)
• Cambridge University Library (Ca)
• Trinity College Library, Dublin (Tr)
Programme of Work

[Maureen Pennock, iPRES 2014]
Approaches to preservation planning

• Digital preservation is business as usual

• Principles for developing a preservation planning capacity:
  – Incremental (no need to do everything at once)
  – Practical (e.g. focused on institutional requirements)
  – Flexible
    • No one-size-fits-all solution (although needs to operate at scale)
    • Needs to cover a wide range of content types
    • Takes into account the whole life-cycle
    • Other stakeholders (e.g. legal deposit libraries)
  – Evidence-based (e.g., assessment-based activities)
Collection profiling (1)

• What content do we have and what is important about it?
  – Digital Asset Register

• Collection profiling:
  – Documenting key knowledge about the Library’s high-level digital collections
  – Considering preservation requirements / preservation intent for these collections
  – A mechanism for liaising with curators and collection specialists
Collection profiling (2)

• Taxonomy of high-level collections:
  – No standard list of content types
  – Various lists are available on the Library website and catalogue, but they are inconsistent
  – Needed a pragmatic starting point for collection profiling:
    • Developed a new taxonomy based on the Library’s existing digital asset register
    • Attempted to identify logical groupings, e.g. ignoring distinctions between digitised and ‘born-digital’ content, where possible
    • It is not perfect (and subject to revision)
Collection profiling (3)

• The profile framework (has evolved over time):
  – Summary
  – What is this format, and what is it within the Library?
  – Acquisition and acquisition formats
  – Preservation Intent
  – Issues
  – Summary table of sub-collections

• A first step in an ongoing process of capturing preservation requirements
Format assessments (1)

• Sustainability assessments of formats
  – Interacts with ongoing discussion on “preservation masters”
  – Less concerned with theoretical “obsolescence,” but aware of institutional factors in maintaining access to content
  – Principles:
    • Based, where possible, on real-world evidence (documented examples)
    • Clarity on purpose and intended audience
    • Not based on numerical scoring measures
Format assessments (2)

• Main criteria
  – Development Status
  – Adoption and Usage
  – Software Support
  – Documentation and Guidance
  – Complexity
  – Embedded or attached Content
  – External Dependencies
  – Legal Issues
  – Technical Protection Mechanisms
  – Other …
Developing preservation planning capacity

• Other tasks:
  – Tool assessment
  – Workflow assessment (considering content lifecycle)
  – Policies
  – Training
Planning for both scale and complexity

- eMSS
- Digital Mapping
- Web archives
- eJournals
- eBooks
- eTheses
- Digitised Sound
- Digitised MSS
- Digitised Newspapers
Summing up

• Preservation Planning in OAIS
  – “The OAIS functional entity which provides the services and functions for monitoring the environment of the OAIS and which provides recommendations and preservation plans to ensure that the information stored in the OAIS remains accessible to, and understandable by, and sufficiently usable by, the Designated Community over the Long Term, even if the original computing environment becomes obsolete.”

• There is a long way to go (but we are getting started)
Thank you
Preservation planning at the Wellcome Library


Dave Thompson
Digital Curator, Wellcome Library
The Wellcome Library has a published Preservation Policy

1. Preservation planning & digital content.
2. Systems & processes

http://2020projectmanagement.com/2013/12/planning-your-project-6-helpful-tips//
Understanding our content

• Built on characterisation reporting out of Preservica.
• Practical hands on investigation.
• Created a spread sheet that is updated monthly.
• Knowing what we have.

<table>
<thead>
<tr>
<th>File format</th>
<th>Version</th>
<th>Format PUID</th>
<th>Frequency</th>
<th>% of BD collection</th>
<th>Approx date of creation</th>
<th>Migration pathway in Preservation?</th>
<th>Supported by vendor?</th>
<th>Rendering application(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypertext Mark-up Language</td>
<td>4.01 fmt/100</td>
<td>20138</td>
<td>51.854</td>
<td>1980s</td>
<td>Yes. To PDF 1.4 &amp; PDFa</td>
<td>CERN / W3C</td>
<td>Web browser / HTML editor</td>
<td></td>
</tr>
<tr>
<td>Acrobat PDF 1.4 - Portable Document Format</td>
<td>1.4 fmt/18</td>
<td>2292</td>
<td>5.906</td>
<td>2001</td>
<td>No, Adobe</td>
<td>Adobe PDF Reader / Various open source PDF readers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>JPEG File Interchange Format</td>
<td>1.01 fmt/43</td>
<td>2244</td>
<td>5.783</td>
<td></td>
<td>Yes. To PNG 1.0 GIF 89a PDF 1.3 PDFa PDF 1.4 &amp; JPEG2000</td>
<td>Yes, Joint Photographic Experts Group</td>
<td>Image rendering / editing software / web browser</td>
<td></td>
</tr>
<tr>
<td>Windows Portable Executable</td>
<td>x-fmt/411</td>
<td>1500</td>
<td>3.865</td>
<td></td>
<td></td>
<td>Microsoft Windows</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tagged Image File Format</td>
<td>fmt/353</td>
<td>1356</td>
<td>3.494</td>
<td></td>
<td>Yes. To PDFa JPEG 1.01 PDF 1.4 PDFa GIF 1989a JPEG 1.02 JPEG2000 &amp; PNG 1.0</td>
<td>Yes, Adobe</td>
<td>Image editing software</td>
<td></td>
</tr>
<tr>
<td>JPEG File Interchange Format</td>
<td>1.02 fmt/44</td>
<td>1157</td>
<td>2.981</td>
<td></td>
<td>Yes. To PNG 1.0 GIF 89a PDF 1.3 PDFa PDF 1.4 &amp; JPEG2000</td>
<td>Yes, Joint Photographic Experts Group</td>
<td>Image rendering/editing software</td>
<td></td>
</tr>
<tr>
<td>Hypertext Mark-up Language</td>
<td>fmt/96</td>
<td>1230</td>
<td>3.170</td>
<td></td>
<td>Yes. To PDF 1.4 &amp; PDFa</td>
<td>Yes, W3C, but depends on version</td>
<td>Web browser / HTML editing application</td>
<td></td>
</tr>
<tr>
<td>Plain Text File</td>
<td>x-fmt/111</td>
<td>1035</td>
<td>2.667</td>
<td></td>
<td></td>
<td>No</td>
<td>Any word processor / text editing tool</td>
<td></td>
</tr>
<tr>
<td>Microsoft Word for Macintosh Document</td>
<td>5 x-fmt/65</td>
<td>661</td>
<td>1.703</td>
<td>1991</td>
<td>Depends on version &amp; network settings, Microsoft</td>
<td>Microsoft Word / Word free viewer / Libre Office / Open Office</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acrobat PDF 1.3 - Portable Document Format</td>
<td>1.3 fmt/17</td>
<td>653</td>
<td>1.683</td>
<td>2000</td>
<td>Yes, To XML 1.0 &amp; PDFa</td>
<td>No, Adobe</td>
<td>Adobe PDF Reader Various open source PDF readers</td>
<td></td>
</tr>
<tr>
<td>Acrobat PDF 1.0 - Portable Document</td>
<td>1 fmt/14</td>
<td>635</td>
<td>1.635</td>
<td>Mid 1990s</td>
<td>Yes, To XML 1.0 PDF 1.3</td>
<td>No, Adobe</td>
<td>Adobe PDF Reader</td>
<td></td>
</tr>
<tr>
<td>File format</td>
<td>Version</td>
<td>Format PUID</td>
<td>Frequency</td>
<td>% of BD collection</td>
<td>Approx date of creation</td>
<td>Migration pathway in Preservica?</td>
<td>Supported by vendor?</td>
<td>Rendering application[s]</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>---------</td>
<td>-------------</td>
<td>-----------</td>
<td>--------------------</td>
<td>-------------------------</td>
<td>---------------------------------</td>
<td>----------------------</td>
<td>-------------------------</td>
</tr>
<tr>
<td>Hypertext Mark-up Language</td>
<td>4.01</td>
<td>fmt/100</td>
<td>20138</td>
<td>51.834</td>
<td>1980s</td>
<td>Yes. To PDF 1.4 &amp; PDFa</td>
<td>CERN / W3C</td>
<td>Web browser / HTML editor</td>
</tr>
<tr>
<td>Acrobat PDF 1.4 - Portable Document Format</td>
<td>1.4</td>
<td>fmt/18</td>
<td>2292</td>
<td>5.906</td>
<td>2001</td>
<td>No, Adobe</td>
<td>Adobe PDF Reader / Various open source PDF readers</td>
<td></td>
</tr>
<tr>
<td>JPEG File Interchange Format</td>
<td>1.01</td>
<td>fmt/43</td>
<td>274</td>
<td>5.783</td>
<td></td>
<td>Experts Group for Image Data Software / Web browser</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Windows</td>
<td></td>
<td></td>
<td>1500</td>
<td>3.865</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tagged PDF</td>
<td>1.02</td>
<td>fmt/44</td>
<td>1157</td>
<td>2.981</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JPEG File Interchange Format</td>
<td>1.02</td>
<td>fmt/44</td>
<td>1157</td>
<td>2.981</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypertext Mark-up Language</td>
<td>fmt/96</td>
<td></td>
<td>1230</td>
<td>3.170</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plain Text File</td>
<td>x-fmt/111</td>
<td></td>
<td>1036</td>
<td>2.667</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Microsoft Word for Macintosh Document</td>
<td>5</td>
<td>x-fmt/65</td>
<td>661</td>
<td>1.703</td>
<td>1991</td>
<td>Depends on version &amp; settings, Data viewer open source PDF readers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acrobat PDF 1.3 - Portable Document Format</td>
<td>1.3</td>
<td>fmt/17</td>
<td>653</td>
<td>1.683</td>
<td></td>
<td></td>
<td>Adobe PDF Reader / Various open source PDF readers</td>
<td></td>
</tr>
<tr>
<td>Acrobat PDF 1.0 - Portable Document</td>
<td>1</td>
<td>fmt/14</td>
<td>635</td>
<td>1.635</td>
<td>Mid 1990s</td>
<td>Yes. To XML 1.0 PDF 1.3</td>
<td>No, Adobe</td>
<td>Adobe PDF Reader</td>
</tr>
</tbody>
</table>

- **How many?**
- **Percentage of our collection**
- **Pink indicates falling numbers**
- **Format**
Tells us…scope of our material

- Single dynamic point of information about our collection.
- Supports decision making around interventions.
- Shows us what is exceptional.
- Supports collection decisions.

http://blog.optiontiger.com/importance-of-technical-analysis/
Further information

- Preservica report a good start.
- Added % of total collection for given fmt.
- Sources of further information.
- Indicators of growth/contraction.

http://www.evanravitz.com/doctor/
But it’s not only about content

Not yet addressed;

• Need to consider systems, processes, IT infrastructure.
• How we process/manage content. Workflows etc.
• How we secure data.
• How we maintain infrastructure.
• Organisational stuff.
Not as bad as we thought

• Have a finite set of tools, set of systems.
• Tools mostly workflow based, predictable & consistent.
• Clear set of processes depending on source/type of content.
• Clear set of roles/responsibilities.
Wellcome Library architecture – system view

Key:
1 = Library catalogues
2 = Good external services
3 = Good
4 = SCR
5 = network storage
6 = Search layer
7 = Player
8 = SIP Creator

Movement of objects
Movement of metadata
System communications

FTP
Delivers the file
Requests files
Requests files
Requests files
Requests files

Created PDF

Welcome Trust network storage

Welcome Trust, May 2011

wellcome trust
Sources of digitised content

Project Managers

- In-house
- Institutions
- Grey literature
- Contractors
- Harvesting

Project Managers / Ingest Officer

- TIFF or JP2
- TIFF or JP2 HD & ftp
- PDF
- TIFF or JP2
- Auto harvesting of JP2 & DMD

Ingest Officer / Digital Curator

- Normalises TIFF to JP2
- Jpylyzer validates JP2
- Auto harvesting of JP2 & DMD

Goobi (METS/OCR)

Preservica

Snagging

Snagging
Standardised processes

Goobi.

Preservica.
What's missing?

- System/workflow documentation.
- Image processing workflow documentation.
- Clear alignment with organisational intent.
- Understanding the environment we have built.
So, to wrap up…

- Very early days.
- Starting by understanding our content.
- Actually know more than we thought.
- Need to better (i.e. more) documentation of environment.
Thank you

Questions now, questions later…?

Dave Thompson, Digital Curator
Wellcome Library

d.thompson@wellcome.ac.uk - @d_n_t

http://wellcomelibrary.org/
Context and Goals
Context
Goals

1. Survey existing best practice
2. Define framework
3. Review
4. Put into action
5. Evaluate and evolve
Existing Best Practice

1) British Library: Digital Preservation Strategy

2) OCLC Research: Preservation Health Check

3) NLNZ: "Digital Preservation in Capable Hands"
Findings

Digital Preservation Strategy

| REASONS |
- Strategy revised and released in 2013
- Contemporary strategy
- Aims differ from Parliamentary Archives
- BL transitioning digital preservation work into BAU

| RESULTS |
- Evidence led approaches
- Embed PP into corporate risk management
- Implement end to end workflows
- Consider sustainability in all system exercises
Findings

Preservation Health Check

| REASONS |
• New approach using metadata for monitoring
• PREMIS as an evidence base
• Threat assessment based on Simple Property-Oriented Threat (SPOT) model

| RESULTS |
• Is such an approach is viable?
• Sensors, thresholds, and triggers
• Potential for automated assessment
• Work in progress
<table>
<thead>
<tr>
<th>REASONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Risk assessment at NLNZ</td>
</tr>
<tr>
<td>• Challenges assumptions</td>
</tr>
<tr>
<td>• Reasoning behind risk assessment approach</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RESULTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Form of active, managed preservation</td>
</tr>
<tr>
<td>• Core ability to render and level of comfort</td>
</tr>
<tr>
<td>• Description of the universe of rendering possibilities</td>
</tr>
<tr>
<td>• Risk is situational</td>
</tr>
</tbody>
</table>
Defining a Framework
Risk Register

CORPORATE FRAMEWORK

1. Clarify Objectives
2. Identify Risks
3. Assess Risks
4. Address Risks
5. Review and Report

RISK HEAT MAP

Impact
1. INsignificant
   1. RARE
   2. UNLIKELY
   3. POSSIBLE
   4. LIKELY
   5. ALMOST CERTAIN

Likelihood
1. MAJOR
2. MODERATE
3. MINOR
4. CATAstroPHIC

Outstanding Questions*

*lots, but not insolvable
1) How can we maximise benefit of existing practice?

2) What threats are we trying to mitigate?

3) How can risk based approaches result in planning actions?

4) How can planning actions be quantified in terms of cost?

5) How can we benchmark and improve?
Enough talk!
Let’s get to work…
1) How does preservation planning fit in with the bigger picture?

2) How do I know what I’m doing is actually preservation planning?

3) How can I tell if the preservation planning activity is effective?

4) How can others – my organisation – know when I’m doing preservation planning?