Open Community Approaches to Digital Preservation

GREY LITERATURE AND DIGITAL PRESERVATION: STANDARDS AND PRACTICE
AMSTERDAM, NOVEMBER 30, 2015

ROSS KING, OPF CHAIRMAN

http://openpreservation.org/
Open Community Approaches to Digital Preservation

THE OPEN PRESERVATION FOUNDATION
Non-profit membership organisation
Founded in 2010

http://openpreservation.org/about/members/
Vision

*Shared solutions for effective and efficient digital preservation*

[http://openpreservation.org/about/organisation/vision-mission-strategy/]
Digital Preservation Community Space

Based on H. Balk (2013) Centres of Competence and Virtual Centres of Excellence

Open Preservation Foundation

Digital Preservation Community Space

DCC/DPC (UK)
NCDD (NL)
NESTOR (DE)
NDIPP (US)
NSLA (AU/NZ)
DPN (US)
JISC (UK)

EC
UNESCO
DLM/RDA/...
OPF/IMPACT/PRESTO/...

Political
Business
Technical
National
EU
International
Vision

*Shared solutions for effective and efficient digital preservation*

Vision

*Shared solutions for effective and efficient digital preservation*

http://openpreservation.org/about/organisation/vision-mission-strategy/
Vision

*Shared solutions for effective and efficient digital preservation*

Mission

*The Open Preservation Foundation sustains knowledge and technology for the long-term management of digital cultural heritage. We provide our members with reliable solutions to the challenges of digital preservation.*

Vision

*Shared solutions for effective and efficient digital preservation*

Mission

*The Open Preservation Foundation *sustains* knowledge and technology for the **long-term management** of digital cultural heritage. We provide our members with **reliable solutions** to the challenges of digital preservation.*

http://openpreservation.org/about/organisation/vision-mission-strategy/
http://openpreservation.org/about/organisation/vision-mission-strategy/
Technology: software

• Open-source stewardship
  – shared infrastructure (hosting, test-driven development)
  – maturity model (influenced by Apache, ISO 25010)
  – products (OPF-Production), incubator (OPF-Labs),
    ‘retirement’ or end-of-life (OPF-Attic)

• Long-term sustainability
  – consistency (hosting and packaging)
  – consolidation (shared road-mapping)
  – effectiveness (testing and evaluation)
  – efficiency (enabling collaboration)

http://openpreservation.org/technology/
Technology: software maturity

• Quality
  – Ensuring that software products conform to standards-driven quality assurance

• Visibility
  – Providing integrated outreach to multiple audiences to maximise discoverability

• Open licensing
  – Using open licenses to encourage the adoption and reuse of source code and software

• Training
  – Enabling skills development to further institutional adoption and capacity building

• Community integration
  – Integrating into commercial and non-commercial products and services

http://openpreservation.org/technology/
## Technology: software quality

<table>
<thead>
<tr>
<th></th>
<th>OPF-Production</th>
<th>OPF-Labs</th>
<th>OPF-Attic</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Community</strong></td>
<td>Active bug reports/fixes</td>
<td>Active bug reports/fixes</td>
<td>[no expectation]</td>
</tr>
<tr>
<td></td>
<td>Named maintainer</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Contribution guidelines</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Product roadmap</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Documentation</strong></td>
<td>Source code comments</td>
<td>Source code comments</td>
<td>Statement describing why the software has</td>
</tr>
<tr>
<td></td>
<td>Technical documentation</td>
<td>Technical documentation</td>
<td>been retired (by original developer or</td>
</tr>
<tr>
<td></td>
<td>Installation manuals</td>
<td></td>
<td>software curator)</td>
</tr>
<tr>
<td></td>
<td>User documentation</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Microsite</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Code quality</strong></td>
<td>Test coverage 50% &gt; 80+%</td>
<td>Test coverage c.50%</td>
<td>[no expectation]</td>
</tr>
<tr>
<td></td>
<td>Continuous integration</td>
<td>Continuous integration</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Integration tests</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Publishing test results</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Deployment</strong></td>
<td>Build from source</td>
<td>Build from source</td>
<td>[no expectation]</td>
</tr>
<tr>
<td></td>
<td>Defined platforms</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Automated packaging (e.g. DEB/RPM/EXE)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(Signed packages)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Support</strong></td>
<td>Online demonstrator</td>
<td>[no expectation]</td>
<td>[no expectation]</td>
</tr>
<tr>
<td></td>
<td>Virtual machine image</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Training materials</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Licensing</strong></td>
<td>Defined open source licence</td>
<td>Defined open source licence</td>
<td>Defined open source license</td>
</tr>
</tbody>
</table>

[github.com/openpreservation](http://openpreservation.org/technology/principles/)
Technology: software portfolio

OPF-Production

• FIDO (format identification)
• JHOVE (format validation)
• JHOVE2 (format validation)
• Jpylyzer (JP2K validation)
• Matchbox (near-duplicate image detection)
• xCorrSound (audio analysis)

[hosted as a service (by us or others)]

• PLATO (preservation planning)
• SCOUT (preservation watch)

OPF-Labs

• Bitwiser (bit analysis)
• Flint (format policy)
• HaWarp (web archive processing)
• Pagelyzer (web archive QA)
• Simulator (scalability)
• Strender (rendering analysis)

[under negotiation]

• Hardware Extractors (TIMBUS)

[contributed to]

• BitCurator (digital forensics)
• FITS (format characterisation)
• c3po (collection profiling)

http://openpreservation.org/technology/products/
Technology Highlight: JHOVE

• JHOVE is one of the most prevalent digital preservation tools used in production
• OPF has recently taken over stewardship of the JHOVE project
  – First steps: stabilization and “mavenization” of the code.
  – New release
  – Next steps: community development roadmap

• As a result, there is a new type of OPF member: the Software Supporter

See http://openpreservation.org/about/join/software-supporters/
Technology: services

• Atlas of Digital Damages – visual examples of failures, e.g. corrupt files
• Catalogue of Policy Elements – guidance, procedures, control
• Datasets, Issues, Solutions – collection context, user requirements, solutions
• File Format Risk Registry – format-specific risks, recommendations and tools
• Tools Registry (COPTR) – reference for >350 digital preservation tools
• Q&A site – question and answer for digital preservation issues

http://openpreservation.org/technology/services/
Knowledge: interest groups

- Community hubs
  - emulation
  - document formats
  - (technical registries)

- Knowledge exchange
  - case studies (blogs)
  - requirements (wiki)

- Technology development
  - evaluation
  - roadmapping

http://openpreservation.org/knowledge/interest-groups/

OPEN PRESERVATION FOUNDATION
Knowledge: events

• Virtual (monthly webinars)
• Face-to-face (c.3 per year)

- Digital Forensics and Emulation
- BitCurator Access: New forms of access to born-digital data
- Long-term preservation with dArcheo
- Preserving PDF: Identify, Validate, Repair
- Executive Seminar: Managing Digital Preservation
- Preserving your preservation tools
- Digital preservation policy
- Tools for uncovering risks in large repositories
- Scalable environments for file format identification and characterisation
- Evidence based preservation planning
- SCAPE tools and services (several)
- Scalable digital preservation (Hadoop)
- From the toolkit: JHOVE2
- Securing funding for digital preservation
- Preservation planning
- Digital library development and practice at LSE
- Application infrastructure for action services (workflow automation)
- Forensics tools and methods (disk images and BitCurator)
- Emulation-as-a-service (CD-ROM ingest and scientific process)
- Content profiling (C3PO)
- Software development and collaborative best practices
- Command-line digital preservation tools
- Preservation health check (using PREMIS)
- Understand the value of preservation data
- Practical tools and approaches to preservation systems
- Database archiving (SIARD, MIXED, RODA)

http://openpreservation.org/events/
Knowledge: training

• Online
  – software development
  – SCAPE training (led by OPF)

• Person-to-person
  – new staff inductions

[forthcoming…]

• ‘How-to’s
  – online demonstrators
  – virtual machines

http://openpreservation.org/knowledge/training/
Advocacy and Alliances

Projects

• PLANETS (EU, FP6)
• SCAPE (EU, FP7)
• Preservation Health Check (OCLC)
• SPRUCE, AQuA (UK, JISC)

[advisory/collaborative]

• E-Ark (EU, ICT-PSP)
• TIMBUS (EU, FP7)

Partnerships

• Digital Preservation Coalition (UK)
• IMPACT CoC (text digitisation)
• PDF Association (PDF standards)

http://openpreservation.org/about/projects/
Advocacy and Alliances Highlight: veraPDF

• veraPDF consortium
  – funded by PREFORMA (EU, FP7/PCP)

• “Definitive” PDF/A Validator
  – open source (GPLv3+/MPLv2+)
  – ISO committee liaison
  – objective frame of reference (corpora)
  – industry adoption (PDF software vendors)

• Policy requirements of memory institutions
  – Policy Profile Registry for sharing best practice

• Integration into existing tools and repositories

http://openpreservation.org/about/projects/verapdf/
Grey Literature and the OPF

THE DDHAPT TOOL
Grey Literature and The British Library

• Grey literature is an important source of information for researchers and the LDLs have traditionally collected and provided access to them in print form. However, these documents are increasingly published on the web, sometimes with no printed counterpart, and so we require a new way of archiving them.

• Since April 2013 the British Library and five other UK Legal Deposit Libraries (LDLs) have collected and provided access to the UK’s digital publications under **Non-Print Legal Deposit (NPLD)** Regulations.

• Digital documents, such as official reports, policy papers, reviews and newsletters published by government, professional, business and university bodies, collectively referred as “grey literature”, are in scope for Non-Print Legal Deposit.
Grey Literature and The British Library

• The British Library has transitioned from a moderate-scale permission-based web archiving operation to large-scale crawling of the UK web, in order to implement NPLD. This has fundamentally changed archival workflows, and has led to the development of new tools to support the domain crawl and drive additional crawling activities (e.g. more frequent crawls of important sites).

• The current workflows for collecting and processing digital documents are time consuming, involve a lot of manual handling, and lead to an unnecessary duplication of effort. It is considered beneficial to streamline and automate these as much as possible, taking advantage of web archiving, where the LDLs already harvest 4 million plus UK websites at least once a year, many of which contain such informally published documents.

→ Digital Documents Harvesting and Processing Tool
DDHAPT Data Flow

“OPERA” Application

DDHAPT Application

Strategic Ingest

Service Portals

Internet

BL DLS

DDHAPT Application

UK WA W3ACT
DDHAPT Data Flow

“OPERA” Application

Targets

DDHAPT Application

SIPs

PDFs

Strategic Ingest

Service Portals

Web Resources

Internet

WARC Files

BL DLS
DDHAPT Advantages

- **Open Source**
  - Deployed to RedHat/Centos6.5 Linux

- **Supports multi-user management of Targets and Documents**
  - Definition of journals and publications
  - Definition of license conditions
  - Annotate based on controlled vocabularies for subjects and collections
  - Supports rapid manual metadata extraction
  - Duplicate and near-duplicate detection

- **Automates numerous aspects for increased efficiency**
  - Takes advantage of automated crawls
  - Creates permanent identifiers and SIPs

See [https://github.com/ukwa/w3act](https://github.com/ukwa/w3act)
# DDHAPT Targets

## Watched Targets

<table>
<thead>
<tr>
<th>Title</th>
<th>URL</th>
<th>Curator</th>
<th>Documents</th>
<th>Crawl Date</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIC Stats</td>
<td><a href="http://www.theaic.co.uk/uk/statistics/aic-stats">http://www.theaic.co.uk/uk/statistics/aic-stats</a></td>
<td>Stewart Robertson</td>
<td>301</td>
<td>88 days ago</td>
<td>Live Crawl</td>
</tr>
<tr>
<td>Alcohol Alert</td>
<td><a href="http://www.iais.org.uk/what-we-do/publication-archive/Alcohol-Alert.aspx">http://www.iais.org.uk/what-we-do/publication-archive/Alcohol-Alert.aspx</a></td>
<td>Stewart Robertson</td>
<td>482</td>
<td>93 days ago</td>
<td>Live Crawl</td>
</tr>
<tr>
<td>Big Innovation Centre Publications</td>
<td><a href="http://www.bijnovationscentre.com/">http://www.bijnovationscentre.com/</a></td>
<td>Stewart Robertson</td>
<td>0</td>
<td>days ago</td>
<td>Live Crawl</td>
</tr>
<tr>
<td>Cambridge Institute for Manufacturing Review</td>
<td><a href="http://www.ims.org.com.ac.uk/research/ims-review/">http://www.ims.org.com.ac.uk/research/ims-review/</a></td>
<td>Sally Habel</td>
<td>4</td>
<td>6 days ago</td>
<td>Live Crawl</td>
</tr>
<tr>
<td>Department for Transport Publications</td>
<td><a href="https://www.gov.uk/government/publications?department%5B%5D=department-for-transport">https://www.gov.uk/government/publications?department%5B%5D=department-for-transport</a></td>
<td>Janine Gronshaw</td>
<td>2255</td>
<td>2 days ago</td>
<td>Live Crawl</td>
</tr>
<tr>
<td>Diabetic Hypoglycaemia</td>
<td><a href="http://www.hypogly.com/">http://www.hypogly.com/</a></td>
<td>Stewart Robertson</td>
<td>3029</td>
<td>112 days ago</td>
<td>Live Crawl</td>
</tr>
<tr>
<td>High Pay Centre</td>
<td><a href="http://highpaycentre.org/">http://highpaycentre.org/</a></td>
<td>Stewart Robertson</td>
<td>48</td>
<td>96 days ago</td>
<td>Live Crawl</td>
</tr>
</tbody>
</table>
### DDHAPT Targets

#### Watched Targets

<table>
<thead>
<tr>
<th>Title</th>
<th>URL</th>
<th>Curator</th>
<th>Documents</th>
<th>Crawl Date</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>International Alert Publications</td>
<td><a href="http://international-alert.org/resources/publications">http://international-alert.org/resources/publications</a></td>
<td>Jennie Grimshaw</td>
<td>0</td>
<td>days ago</td>
<td>Live Crawl</td>
</tr>
</tbody>
</table>
DDHAPT Targets - Detail View
## DDHAPT Documents

<table>
<thead>
<tr>
<th>Title</th>
<th>Landing Page</th>
<th>Document Name</th>
<th>Crawl Date</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACMD advice about Northern Ireland’s continued maintenance of the Addicts Index</td>
<td><a href="https://www.gov.uk/government/publications/acmd-advice-about-northern-irelands-continued-maintenance-of-the-addicts-index">https://www.gov.uk/government/publications/acmd-advice-about-northern-irelands-continued-maintenance-of-the-addicts-index</a></td>
<td>...ACMD_letter_es__added_addoe__final.pdf</td>
<td>95 days ago</td>
<td>ignore</td>
</tr>
<tr>
<td>ACMD advice on independent prescribing of controlled drugs</td>
<td><a href="https://www.gov.uk/government/publications/acmd-advice-on-independent-prescribing-of-controlled-drugs">https://www.gov.uk/government/publications/acmd-advice-on-independent-prescribing-of-controlled-drugs</a></td>
<td>...144117/6epend-prescribing-advice.pdf</td>
<td>126 days ago</td>
<td>ignore</td>
</tr>
<tr>
<td>ACMD advice on ‘Ivory Wave’</td>
<td><a href="https://www.gov.uk/government/publications/acmd-advice-on-ivory-wave">https://www.gov.uk/government/publications/acmd-advice-on-ivory-wave</a></td>
<td>...Ivory_wave.pdf</td>
<td>126 days ago</td>
<td>ignore</td>
</tr>
<tr>
<td>ACMD advice on lisdexamfetamine and 2-drugs</td>
<td><a href="https://www.gov.uk/government/publications/acmd-advice-on-lisdexamfetamine-and-2-drugs">https://www.gov.uk/government/publications/acmd-advice-on-lisdexamfetamine-and-2-drugs</a></td>
<td>...data/file/52785/ACMD_advice_7_drugs.pdf</td>
<td>129 days ago</td>
<td>ignore</td>
</tr>
<tr>
<td>ACMD advice on lisdexamfetamine and 2-drugs</td>
<td><a href="https://www.gov.uk/government/publications/acmd-advice-on-lisdexamfetamine-and-2-drugs">https://www.gov.uk/government/publications/acmd-advice-on-lisdexamfetamine-and-2-drugs</a></td>
<td>...data/file/52785/ACMD_advice_7_drugs.pdf</td>
<td>129 days ago</td>
<td>ignore</td>
</tr>
<tr>
<td>ACMD advice on phencyclidine</td>
<td><a href="https://www.gov.uk/government/publications/acmd-advice-on-phencyclidine">https://www.gov.uk/government/publications/acmd-advice-on-phencyclidine</a></td>
<td>...file/191642/advice-on-phencyclidine.pdf</td>
<td>127 days ago</td>
<td>ignore</td>
</tr>
<tr>
<td>ACMD advice on salbutamol</td>
<td><a href="https://www.gov.uk/government/publications/acmd-advice-on-salbutamol">https://www.gov.uk/government/publications/acmd-advice-on-salbutamol</a></td>
<td>...ent_data/file/144115/advice-salbutamol.pdf</td>
<td>125 days ago</td>
<td>ignore</td>
</tr>
<tr>
<td>ACMD advice on steroids to inhalers</td>
<td><a href="https://www.gov.uk/government/publications/acmd-advice-on-steroids-to-inhalers">https://www.gov.uk/government/publications/acmd-advice-on-steroids-to-inhalers</a></td>
<td>...data/file/189134/advice-on-steroids-1.pdf</td>
<td>126 days ago</td>
<td>ignore</td>
</tr>
<tr>
<td>ACMD advice on the scheduling of GHB</td>
<td><a href="https://www.gov.uk/government/publications/acmd-advice-on-the-scheduling-of-gbh">https://www.gov.uk/government/publications/acmd-advice-on-the-scheduling-of-gbh</a></td>
<td>...ent_data/file/045191/ACMD_GHB_Advice.pdf</td>
<td>130 days ago</td>
<td>ignore</td>
</tr>
</tbody>
</table>
DDHAPT Metadata Extraction

**Documents > 2012-University-of-Westminster-Financial-statements**

<table>
<thead>
<tr>
<th>View</th>
<th>Edit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>2012-University-of-Westminster-Financial-statements</td>
</tr>
<tr>
<td>DCI (Digital Object Identifier)</td>
<td></td>
</tr>
<tr>
<td>Publication Date</td>
<td></td>
</tr>
<tr>
<td>Year of Publication</td>
<td></td>
</tr>
<tr>
<td>Filename</td>
<td>2012-University-of-Westminster-Financial-statements.pdf</td>
</tr>
<tr>
<td>Priority Cataloguing</td>
<td></td>
</tr>
<tr>
<td>Services</td>
<td>Business, STM</td>
</tr>
<tr>
<td>Subjects</td>
<td>Adult education, Advertising, Alternative medicine, Anthropology, Archives, Beverages</td>
</tr>
<tr>
<td>Submission Type</td>
<td></td>
</tr>
</tbody>
</table>
DDHAPT Metadata Extraction
DDHAPT Metadata Extraction
DDHAPT Document Types
DDHAPT Documents
DDHAPT SIP Export
# DDHAPT Subjects

![DDHAPT Subjects Page](image)

## FAST Subjects

Enter FAST Subject Name

<table>
<thead>
<tr>
<th>Name</th>
<th>FAST ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adult education</td>
<td>f000797275</td>
</tr>
<tr>
<td>Advertising</td>
<td>f000797631</td>
</tr>
<tr>
<td>Alternative medicine</td>
<td>f000801653</td>
</tr>
<tr>
<td>Anthropology</td>
<td>f000810196</td>
</tr>
<tr>
<td>Archives</td>
<td>f000814030</td>
</tr>
<tr>
<td>Beverages</td>
<td>f000831042</td>
</tr>
<tr>
<td>Business</td>
<td>f000842202</td>
</tr>
<tr>
<td>Business enterprises</td>
<td>f000842519</td>
</tr>
<tr>
<td>Business planning</td>
<td>f000842819</td>
</tr>
<tr>
<td>Civilization</td>
<td>f000963699</td>
</tr>
<tr>
<td>Civil Rights</td>
<td>f000866327</td>
</tr>
<tr>
<td>Communities</td>
<td>f001420062</td>
</tr>
<tr>
<td>Continuing education</td>
<td>f000976706</td>
</tr>
<tr>
<td>Corrections</td>
<td>f000880500</td>
</tr>
<tr>
<td>Crime</td>
<td>f000880684</td>
</tr>
<tr>
<td>Criminal justice, administration of</td>
<td>f000893246</td>
</tr>
<tr>
<td>Criminology</td>
<td>f000893039</td>
</tr>
<tr>
<td>Culture</td>
<td>f000880559</td>
</tr>
<tr>
<td>Diplomacy</td>
<td>f000894198</td>
</tr>
<tr>
<td>Diseases</td>
<td>f000891516</td>
</tr>
</tbody>
</table>
DDHAPT Organisations

Users & Agencies > Organisations

Enter name to search for organisation

Title

- Austrian Institute of Technology
- Bodleian Library
- British Stand-up Comedy Archive at the University of Kent
- Cambridge University Library
- Queen's University Belfast
- The British Library
- The National Library of Scotland
- The National Library of Wales
- Trinity College Dublin
Summary

• The Open Preservation Foundation suggests a successful model for open community approaches to Digital Preservation.
• In these times of lean funding, synergies and economies of scale offered by community approaches are more important than ever.
• The OPF is providing software stewardship for some of the most important open source tools, in particular JHOVE and veraPDF.
• OPF members are producing useful open source tools for the preservation community.
• DDHAPT in particular could be a useful tool for the Greynet community.
• AIT offers research and development services in this area.
• Please consider becoming an OPF Member or Software Supporter!
Thank you for your attention!

Questions?

http://openpreservation.org
http://www.ait.ac.at