**REPORT TO THE LIBRARY BOARD**

**MEETING DATE: NOVEMBER 19, 2008**

<table>
<thead>
<tr>
<th>Session:</th>
<th>Public</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject:</td>
<td>Internet Service Research Report: Open Source Filtering</td>
</tr>
<tr>
<td>Prepared By:</td>
<td>Margaret Mitchell, Tom Travers, David Mitchell</td>
</tr>
<tr>
<td>Presented By:</td>
<td>Tom Travers, Margaret Mitchell</td>
</tr>
<tr>
<td>Purpose of Report:</td>
<td>For Receipt and Information Only ✓</td>
</tr>
</tbody>
</table>

---

**Recommendation:**

It is recommended that this report, including Appendices 1 and 2 be received.

**BACKGROUND & REVIEW**

At its June 2008 meeting, the Library Board received the Public Computer Use and Internet Access Policy Update (L08/33.1). Following discussion, it was moved that staff provide the Library Board with information about "open source filtering", including resource and cost requirements.

This information is provided in
- Appendix 1: Report on Open Source Filtering
- Appendix 2: Open Source Feature Comparison List
APPENDIX 1: REPORT ON OPEN SOURCE FILTERING

Prepared by: David Mitchell, Tom Travers

DEFINITIONS

Open Source software is a development methodology which provides access to a product's source code and design. It is made freely available through a number of licensing models.

Closed Source or proprietary software is computer software on which the producer has set restrictions on use, private modification, copying, or republishing. The internal mechanisms of how the product works are not available to the user unless the software developer chooses to do so.

PRODUCTS REVIEWED

Two open source products were selected for review. These products were compared to the existing Netsweeper product currently in use at London Public Library (LPL). These products were chosen for review because of:

- Existing implementation (NetSweeper)
- Mature, open source product (DansGuardian)
- Commercial implementation of a mature, open source product (SmoothGuardian).

While there are a number of other products (such as squidGuard, POESIA) on the market, initial analysis of feature set indicated they would not be suitable for comparison or ready for production environment.

Netsweeper

LPL's current content filtering system is a closed-source content filtering system, utilizing a proprietary algorithm to classify over a billion URLs. The library chooses which categories it will implement on which type of computers. The library may also specify its own URLs to allow/deny or reclassify. If a customer is blocked from a site that it is felt to be incorrectly classified Netsweeper presents a NetAlert interface for reporting to the vendor and library staff. Netsweeper runs on the open-source Linux operating system and utilizes a number of open-source applications, but its core is not open. It is both 1CIPA1 and 2BECTA2 certified.

Dans Guardian

DansGuardian is an open-source content filtering system. Primarily using a combination of URL blacklists and phrase weighted per page blocking, it also supports 3PICS3 and file type filtration. Phrase lists are user customizable as are the order and type of filtration. It also supports a NetAlert-style customer reporting feature. While it is an actively developed and mature product it does not have any guaranteed software support. The author of the software requests a onetime download license for commercial use but the product is free otherwise. The author cites that the product can be configured to meet CIPA compliance. Third party implementations have been tuned to BECTA compliance as well.

---

1 CIPA - Children's Internet Protection Act (CIPA) is a federal law enacted by the United States' Congress in December 2000 to address concerns about access to offensive content over the Internet on school and library computers.
2 BECTA - British Educational Communications and Technology Agency known as BECTA is the United Kingdom Government's key partner in the strategic development and delivery of its information and communications technology (ICT) and e-learning strategy for the schools and the learning and skills sectors.
3 PICS - Platform for Internet Content Selection. Specification from the W3C to define the content of your site for your visitors by associating labels (metadata)
**SmoothGuardian**

Smooth Guardian is a commercially supported implementation of DansGuardian built on the SmoothWall firewall platform. It has all the features of DansGuardian with the addition of protocol, time based and anti-virus/malware filtering. It is both CIPA and BECTA certified.

**OPEN SOURCE FEATURE COMPARISON LIST**

A comparison of features is found in Appendix 2.

**ANALYSIS**

**Product Effectiveness: Over or Under Blocking of URLs**

The successfulness of any filtering software product is in its ability to minimize errors in the over or under blocking of URL's. *Under blocking* occurs when filtering software fails to block sites because the software cannot properly analyze it. *Over blocking* occurs when filtering software accidentally blocks legitimate sites that do not fit the filtering category.

The two open source products reviewed do not publish percentage error rates. Netsweeper claims an error rate of +/- 4% in the category of pornography. There is no simple way of evaluating the error rates on any of the products from a technical perspective. No product would claim an error rate of zero.

Potential errors in categorization are monitored through customer and staff feedback. LPL has put in place a number of mechanisms for customers and staff to report potential errors in over and under blocking:

- Net Alerts are sent directly and anonymously from the customer to Netsweeper and monitoring emails are provided to LPL;
- URL's can be reported to staff and sent to Netsweeper directly for review;
- "How Did We Do Today" forms can be submitted to LPL, anonymously or with the name of customer, along with comments; and
- Ability to monitor monthly reports from Netsweeper re category/URL.

**Cost**

Some open source software, such as DansGuardian, is free to acquire. However, there is potentially high staff resource cost incurred when using such product. Initial costs would be human resource costs to implement. The software is not a turnkey solution therefore system maintenance, such as the installation of upgrades, would require more expertise and time. The pre-configured reports to satisfy the library's requirements would likely not exist and would need to be developed. Software changes beyond configuration of the basic software would require contracted development.

SmoothGuardian is a commercially supported implementation of DansGuardian. As with Netsweeper there are annual service fees. Its similar turnkey nature means it would not require resources to maintain. This does however reduce the level of customization possible with DansGuardian. Initial costs would include staff resources to configure to meet the current implementation.
An estimated cost comparison is provided below:

<table>
<thead>
<tr>
<th>Estimated Comparable Costs: Open Source Software Compared With Netsweeper*</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Open Source</strong></td>
</tr>
<tr>
<td>Software costs</td>
</tr>
<tr>
<td>Maintenance, including upgrades</td>
</tr>
<tr>
<td>LPL Staff/Contract Resources:</td>
</tr>
<tr>
<td>Setup, Initial configuration</td>
</tr>
<tr>
<td>Ongoing Maintenance</td>
</tr>
</tbody>
</table>

*Does not include software changes beyond basic configuration, alterations to word lists, algorithms, or technical product assessment.

**Policy Development and Application**

The utilization of open source software requires that the LPL review the software options and define a policy that governs their configuration. The exact nature of what was to be blocked would need to be clearly defined. LPL would need to define the:

- Category or categories;
- Scope of each specific category;
- Content in terms of words, graphics, images, sound, etc.; and
- Context and proximity relationships of words, graphics, images, sound, etc.

The next step would be to create a configuration and to determine if the default configuration meets the definition outlined in the policy. If the default configuration does not meet the requirement of LPL, customization can then be made to this configuration (for example defining the required number of times for a particular word to appear in the content, juxtaposed to what other words, the weighting of the words, etc.). These would all be decisions that would rest with LPL.

If it is determined that the customized configuration is still inadequate LPL would need to determine if the algorithms (the software that works with the defined configuration) need to be customized. Outside resources or co-ordination with official software support may be needed to modify them as LPL has neither the staff expertise nor the financial resources to develop such complex technical specifications.

**CONCLUSION**

The two primary advantages of open source software are: transparency and flexibility. The user can have access to word lists and algorithms and, depending on need and resources, alter those to meet specifically defined needs which are not met by the base product.

Open source software, such as DansGuardian would provide the LPL with the opportunity to access and modify the word list as well as the code for the categorizing algorithms and to modify both according to LPL developed policies. However, modifying the word list requires significant knowledge in areas being targeted and this expertise is not available at the Library. The software and programming expertise required to develop or modify the actual code to create the algorithms is currently not available at the Library. In both cases, expertise would need to be contracted or hired to manage this aspect of the operation. No software upgrades would be available.

As described above, policy development would be required to direct the creation of categories, scope of categories, content context and proximity, and the programming of code. Understanding the impact of this work on the outcome is not an area of expertise that is currently available in the Library. Netsweeper provides these services within its contract.
The evaluation of the effectiveness of the LPL application of open source software to the filtering is evaluated on the same factors as the current Netsweeper product, that is, customers and staff reporting potential errors in over and under blocking and follow up by Library staff. Any further level of technical evaluation of the open source product would require resources far beyond the Library's current capacity.

**SUMMARY**

- While open source software may have potential benefits in terms of accessing the software, LPL does not currently have staff resources or expertise to support the product in a modified form or extended configuration.
- The rate of over blocking or under blocking of an open source is not available or published.
- Open source software at this time has not statistically demonstrated that it provides a better solution than the current Netsweeper filtering software.
- The evaluation methodologies LPL currently uses would continue to be the only way of effectively monitoring this customer service, without significant resources to acquire expertise.
### APPENDIX 2: OPEN SOURCE FEATURE COMPARISON LIST

Please note: *Asterisk in chart means review information in notes column

<table>
<thead>
<tr>
<th>Feature</th>
<th>Netsweeper</th>
<th>SmoothGuardian</th>
<th>DansGuardian</th>
<th>Description</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>IWF URL Filtering</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Internet Watch Foundation, an online policing organization</td>
<td></td>
</tr>
<tr>
<td>Turnkey solution</td>
<td>Y</td>
<td>Y</td>
<td>N*</td>
<td>Aside from customization, product ships in a usable configuration</td>
<td>Virtual Appliance provides turnkey like - requires ongoing maintenance</td>
</tr>
<tr>
<td>Paid Support</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Directly block a site that matches on blacklist regardless of content</td>
<td></td>
</tr>
<tr>
<td>URL Blocking</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Regardless of site classification, if keywords match the objectionable page will be blocked</td>
<td>Y=Search query only N=URL block</td>
</tr>
<tr>
<td>Keyword Blocking</td>
<td>Y*</td>
<td>Y</td>
<td>N</td>
<td>Allow select words and combinations of words to determine filtering</td>
<td>Feature available but proprietary feature</td>
</tr>
<tr>
<td>Keyword Weighting</td>
<td>Y*</td>
<td>Y</td>
<td>Y</td>
<td>Access or levels of access can be determined by time of day</td>
<td></td>
</tr>
<tr>
<td>Other Internet protocols</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Products like Skype, MSN Messenger, BitTorrent clients</td>
<td></td>
</tr>
<tr>
<td>Out of the box Reporting</td>
<td>Y</td>
<td>Y</td>
<td>Y*</td>
<td>Third party reports, customizable</td>
<td></td>
</tr>
<tr>
<td>Custom Reporting</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Unused at this time, but allows a particular user account to have different policies than the default</td>
<td>Not as many authentication options as the other products</td>
</tr>
<tr>
<td>User Authentication</td>
<td>Y</td>
<td>Y</td>
<td>Y*</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>