DESIDOC Journal of Library & Information Technology, Vol. 32, No. 6, November 2012, pp. 493-498

Exposing Library Catalogues to Search Engines

Vinit Kumar

Institute of Library and Information Science, Bundelkhand University, UP-284 128

ABSTRACT

The paper attempts to provide ways to reach out to the users by exploiting present day mighty web search engines. Present day library Web OPACs architecture does not really help search engine robots or crawlers to index the huge library data. By exploiting some of the best practices of information architects and webmasters, libraries can also open their huge data to the search engines and can get listed in the top results to get more visibility. This paper describes the problem of unfriendliness of library OPACs and the reasons behind this. The paper further suggests the use of sitemaps to expose the bibliographic records to search engines. Further by discussing the different options to create, upload, and submit the sitemaps to search engines, the paper moves to list some of the benefits and concludes by giving some future insights in this area.

Keywords: Sitemaps, search engines, OPAC, library websites, visualisation

1. INTRODUCTION

Since the inception of internet technology, the abilities of the internet search engines for finding resources have always marveled. A few terms are typed, wait for some milliseconds and the search engine would return a list of results matching query to pages that are supposed to have the terms one is looking for. This has led to the belief that if something is not available through search engines it does not exist at all. With the passage of time these search engines have become a gateway for starting anything which is existing on the present day internet.

Having said that, it is known that the library catalogues they have not evolved themselves to get indexed by search engines. The library catalogues and other library services are of varied architecture. Some of them use their own search engines based on technologies like Lucene, Solr, etc.[1,2] Some are developed commercially such as, Endeca[3], Primo[4] etc. These search engines perform well but they provide very complex or sometimes non-intuitive interfaces, which actually are not upto the expectations of the present day users accustomed to the above mentioned settings. Apart from this, the data of libraries are hidden behind databases or in such a setting that the robots or crawlers of popular search engines suppose that it is a trap to fool them and they do not index it.

This paper describes the problem of unfriendliness of library OPACs towards the web search engine crawlers and some probable reasons behind this. The paper suggests the use of sitemaps to expose the library's bibliographic records to search engines.

2. BACKGROUND

In 2005, a major study done by OCLC[5] (Online Computer Library Center) revealed that 72 percent respondents used Web search engine and only 30 percent used library catalogue at least once. In terms of beginning the search for information on a particular topic, 84 percent respondents start their search from web search engines while only 1 percent of them start from library. Similarly, 80 percent of respondents said the search engine would be their first choice the next time they need a source for information. This study clearly reveals that the users are more satisfied with the services offered by web search engines.

On the other hand the holdings and other bibliographic details stored inside the library catalogues are not indexed by the web search engines. This requires users to separately search in the library OPACs. For example, a resource which exists in the library is visible to the user after querying on a popular web search engine showing its presence on Amazon but not in library website will imply to the user that the particular resource is not available in

Received 22 December 2011, accepted 08 February 2012, online published 19 November 2012

4.1 What is a Sitemap?

A sitemap in its simplest form is a simple plain-text list of all the URLs from a particular website[9] It may also be written in a recognised syndication format such as RSS, mRSS, or Atom 1.0. The drawback with using the RSS or Atom is that this method provides crawlers with only recently created content. But most often the sitemaps are written in XML following the sitemap XML protocol. The protocol has provision to provide some mandatory information and some optional information[10].

Mandatory elements:

<urlset>

- The document-level element for the Sitemap. The rest of the document after the '<?xml version>' element must be contained in this.

<url>

- Parent element for each entry. The remaining elements are children of this.

<loc>

- Provides the full URL of the page, including the protocol (e.g. http, https)

Additional elements:

<lastmod>

- The date that the file was last modified, in ISO 8601 format. This can display the full date and time or, if desired, may simply be the date in the format YYYY-MM-DD.

<changefreq>

- How frequently the page may change: always, hourly, daily?

<priority>

- The priority of that URL relative to other URLs on the site.

The use of additional elements varies from search engine to search engine their use is not mandatory but it helps crawlers to crawl websites more intelligently.

A sample sitemap:

<?xml version='1.0' encoding='UTF-8'?>

<urlsetxmlns="http://www.sitemaps.org/schemas/sitemap/0.9"

xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"

xsi:schemaLocation="http://www.sitemaps.org/schemas/sitemap/0.9

http://www.sitemaps.org/schemas/sitemap/0.9/sitemap.xsd">

<url>

<loc>http://drtc.isibang.ac.in/</loc>

<lastmod>2010-11-18</lastmod>

<changefreq>daily</changefreq>

<priority>0.8</priority>

</url>

<url>

<loc>http://drtc.isibang.ac.in/DRTC/KT</loc>

<lastmod>2009-12-12</lastmod>

<changefreq>monthly</changefreq>

<priority>0.8</priority>

</url>

</urlset>

4.2 Creation of Sitemaps

There are different ways to create sitemaps. The libraries can create sitemaps essentially by three approaches:

(1) Locally-run, hand-coded routines

This method can prove to be little harder and time taking. It requires the technical knowledge of the sitemap protocol as well as its restrictions. Similarly, the programming knowledge of the language in which the tool is written is also required. Some of the available codes are[11]:

- GSiteCrawler

  http://gsitecrawler.com/

- G-Mapper

  http://www.dbnetsolutions.co.uk/gmapper/

- WebDesignPros Sitemap Generator

  http://www.webdesignpros.ca

(2) Remotely called sitemap generation software/services

This approach is good for only those kind of websites where the content is not obscure as this involves connection from remotely located software or service providers. Otherwise there will be some efforts required to

to exclude and include the areas of the site which the website creator wants the crawler to index. So the best option to submit sitemaps to search engines would be to use robots.txt file to direct to the location of sitemap. The sitemap specifications suggests following format:

sitemap: http://drtc.isibang.ac.in/sitemap.xml.gz

Another option is to submit the sitemap to popular search engines, such as Google, Yahoo, Bing etc. directly. For example, Google provides lot many tools for webmasters to make standard websites through its 'Google's Webmaster Tools' where webmasters can register and can upload the sitemaps and specify the frequency of the Googlebot (Google's Crawler) visits. This requires signing up for Google Account.

Similarly, the present day web search engines provide ping URLs on which the website creator can ping an http request and submit the sitemaps. Ping URLs of some of the search engines are[13]:

Google: http://www.google.com/webmasters/sitemaps/ping?sitemap=

Ask: http://submissions.ask.com/ping?sitemap=

Bing: http://www.bing.com/webmaster/ping.aspx?siteMap=

Yahoo: http://search.yahooapis.com/SiteExplorerService/V1/updateNotifi cation?appid=SitemapWriter&url=

The sitemap URL of the library OPAC should be suffixed after "=" sign.

4.5 Benefits

The present day movement of users to the popular web search engines makes it more important to make library online services search engine-friendly. The benefits of having sitemaps are many for the library bibliographic services.

The display of library records on search engines will yield following benefits:

- Save time of users by giving search feature on a single interface.
- Bring more satisfaction to users about the library services.
- Help in better discovery of library records.
- Increase traffic to the library OPAC which further will provide more visibility.
- Place library OPAC to be on par with other popular destinations on internet.

5. CONCLUSIONS

Use of sitemaps is very popular among the webmaster community to expose websites to search engines. The best practice is to provide a local search engine embedded inside the website as well as making the website Web search engine friendly. Libraries in their long history always focused on the second part but are far behind in developing the strong connection with internet search engines. It is not that library OPACs are not interoperable, the use and Z39.50 protocol and at present SRU/SRW proves that they also tried to bring federated searching on the picture. But these services and protocols were only to communicate between the libraries situated at different locations and not for the present day search engines.

Future work in this area could be to develop web-browser add-on (plugins) to provide 'localised search engines' or library specific search engines for the library catalogues so that the users would search library catalogues inside the search bar in their browser without visiting the OPAC. This can be done by having library specific search engine built using opensearch description specifications[14]. This will further give users an option to install the add-on in their browsers on their first visit to the library website. On the similar grounds another approach could be to tweak the results page of the search engines and to push the hits from the library up in the list. These add-ons can be build starting from scratch or customising some already existing add-on. The need of installing it on the local computer of user will be a drawback that can be resolved by on-campus publicity from the library.

REFERENCES

1. Apache. Apache Lucene-overview. 17 June 2010. http://lucene.apache.org/java/docs/index.html (accessed on 12 November 2010).
2. Apache. Welcome to Solr. Global search market draws more than 100 billion searches per month. 14 July 2010b. http://lucene.apache.org/solr/comScore, 2009 (accessed on 14 November 2010).
3. Endeca. Search and business intelligence software for ecommerce and the enterprise. 2009. http://www.endeca.com/en/home.html (accessed on 14 November 2010).
4. Lewis, N. Implementing Ex Libris's Primo at the University of East Anglia. *Ariadne,* April 2008, **55**. http://www.ariadne.ac.uk/issue55/lewis/
5. OCLC. Perceptions of libraries and information resources: A report to the OCLC membership. OCLC, Ohio, 2005. pp. 12-19.