Case Study - Search Engine Censoring of Sites Using Open Directory Data

Ted Goldsmith
Azinet LLC
March 27, 2006

Search Engine Honesty

Introduction

The Open Directory Project (ODP) (http://dmoz.org/) employs 71,000 volunteer editors to produce a database of edited and reviewed web site listings. (This figure probably represents all the editors that have ever worked on ODP.) Currently the database contains more than 5 million site listings in about 600,000 categories and 31 languages. ODP was originally started as "GnuHoo", a knockoff of the Yahoo directory design. It is now operated by Netscape, a division of AOL. The Open Directory appears under at least one other Netscape domain name at http://newhoo.com/.

ODP editors write the site title and site description to be placed in the directory for each listed site. They also visit the site and verify that the site is about the subject specified for the category in which it appears.

ODP editors are selective and do not accept every site: "While the ODP is comprehensive in scope and coverage, we care a great deal about the quality of the ODP, and pride ourselves on being highly selective. We don't accept all sites, so please don't take it personally should your site not be accepted. Our goal is to make the directory as useful as possible for our users, not to have the directory include all (or even most) of the sites that could possibly be listed or serve as a promotional tool for the entities listed."

The data is available ("open license") for use by any web site. Unlike most directories ODP does not charge for listings: "There is not, nor will there ever be, a cost to submit a site to the directory, and/or to use the directory's data. The Open Directory data is made available for free to anyone who agrees to comply with our free use license."

ODP data is widely used: "The Open Directory powers the core directory services for the Web's largest and most popular search engines and portals, including Netscape Search, AOL Search, Google, Lycos, HotBot, DirectHit, and hundreds of others."

ODP data is a valuable resource. However, use of ODP data has some serious search engine issues described below. As a result, major search engines have completely censored access by their users to many sites using ODP data. This study examines these issues and provides information regarding search engine censoring of sites using ODP data.

ODP data can be used by webmasters in a number of different ways. ODP provides access to their data via free download of two large files in "RDF" (a type of XML)

format. Web sites can write software to periodically read these files, extract all or a selected subset of the data, and construct web pages to be hosted on the site's server. Free or commercial software is available to easily perform these functions to produce a "clone" of the entire ODP directory on a web site server. ODP data can also be extracted from these files to populate a web site's database system in order to power a dynamic web site directory in which pages are generated in real time from data hosted on the web site's server.

Another much easier and cheaper way to use ODP data is to build a "virtual" ODP clone. In this method, the site doesn't actually host the ODP pages or data. Instead, a small software package on the site intercepts page requests (clicks) to the site requesting ODP data. The software then executes an automated page request to ODP's site requesting and downloading the corresponding page to the server. Then the software "scrapes" off the ODP logo and other identifying information and replaces them with the web site's logo, site unique text, and ads before sending the page to the user's browser. This "live" method allows a web site to have a "virtual" copy of ODP's 600,000 page directory without expending any "real" resources (disc space, programming time, maintenance, etc.). Because of the extra communications between the site's server and the ODP server, and because ODP's server is heavily loaded by all the parasitic virtual clones, this method responds more slowly than the first method. During peak times it is often currently unusable.

ODP encourages web site owners to use ODP data including the use of the free software to make essentially identical "clones" of the entire ODP 600,000 page directory and including use of the "live" software for producing maintenance-free and resource-free "virtual" clones of the entire 600,000 page directory.

A review of several free software packages for producing ODP clones describing how they work can be found at: http://www.10-minute-rule.com/index.php?itemid=14 (ODP provides a link to this page.)

An ODP page with links to free and commercial software for building real and virtual ODP clones is found at:

http://dmoz.org/Computers/Internet/Searching/Directories/Open_Directory_Project/Use_of_ODP_Data/Upload_Tools/

Thumbshots (http://thumbshots.org) provides free "thumbnail" screen shot images of sites listed in ODP to "over 1000 sites" to enhance sites using ODP data. They provide their own clone of ODP, with thumbshots, at http://www.thumbshots.net .

Search Engine Issues with Sites Using ODP Data

Search engines have potentially major problems with this situation as follows:

1) Search engines use "link popularity" as a factor in determining how to rank search results. The idea is that links coming from other web sites indicate that a site is "popular" and therefore might be "more important" and of higher "quality" than some other site that

has equal relevance to the search terms. Directories, since they have links to thousands or millions of other sites tend to disrupt the link popularity concept. (Google's *PageRank* system is thought to be less sensitive to this problem.)

2) The complete ODP directory has more than 600,000 pages. There is free or commercial software available that allows site owners to very easily and inexpensively create "clones" of ODP that are essentially exact duplicates with no "value added". Because of the very large number of pages potentially involved, search engines might be investing major resources (disc space, memory, computer time, bandwidth) in indexing and spidering essentially duplicate data, if they index a large number of pages on the clones. (2,000 sites producing 600,000 pages each is *1.2 billion pages*, a significant fraction of the total pages indexed by a major search engine (about 8 billion pages).

However, search engines have means for limiting "depth-of-crawl" (number of pages indexed relative to total number of pages on the site) other than outright manual censoring ("banning"). Our data indicate that search engines often do not index very many pages on ODP clone sites.

3) Google itself publishes a copy of the ODP directory (http://dir.google.com/). Google also indexes its own directory. That is, a Google search can return results that link to Google's copy of the Open Directory. Google might well consider other sites using ODP data to be competition and therefore suppress access to the competing sites via Google Search. Our data show that Google is much more likely to censor sites using ODP data than the other major search engines.

Are search engines therefore systematically censoring (that is deliberately, "manually", blocking, on a site by site basis) their user's access to entire sites using ODP data? We conducted a study to find out.

Method

Open Directory publishes a list of more than 300 "sites using ODP data" at:

http://www.dmoz.org/Computers/Internet/Searching/Directories/Open_Directory_Project/Sites_Using_ODP_Data/full-index.html

This list probably only includes a small fraction of sites using ODP data and does not include the major players. ODP is an edited directory of reviewed sites that does not publish all sites submitted. The published sites are therefore presumably representative of the "better" sites using ODP data. ODP's list also does not include many sites whose owners did not submit their site to be listed, or sites recently submitted and not yet processed by ODP's editors.

We examined the first 40 sites on ODP's list. From the point of view of the study, these sites, having been selected by ODP and ordered alphabetically, were considered a reasonable random sample of all the sites using ODP data. There was no reason to believe that search engines were more or less likely to censor these sites than other sites in the ODP list. If "quality" was a factor in search engine censoring, that is, if search

engines were more likely to censor low quality sites, search engines could be expected to be more likely to censor sites using ODP data that had *not* been included in the ODP "selective" list.

The sites were visited between 3/20/2006 and 3/22/2006.

Alexa (www.alexa.com) measures total site (domain.com) traffic using a free browser toolbar that reports to Alexa on each page visited by a toolbar user. Alexa reports the rank (1 corresponds to highest traffic) of any site. If no toolbar user has visited the site, Alexa reports "no data". In our results listed below, we show a rank of "> 4M" for sites where Alexa indicated no data.

Google also has a free toolbar that reports "PageRank" (PR) of any page that is being viewed in the browser. PageRank is represented by Google to be an "honest" "indicator" of site "quality" that is "automatically calculated". PageRank is displayed as a number between zero (minimum "quality") and ten (maximum "quality"). In our results we indicate the PageRank indicated by the Google toolbar for pages in each site.

All the major search engines allow a search to be performed where results are limited to pages on a particular web site. This is done using a search for "site:hostname.com". Search engines also report the approximate number of pages found in the search engine's index as a result of a search. If a site has been totally censored by a search engine, the "site:" search returns zero or sometimes one page found. Otherwise "site:" returns the approximate number of pages from that site that have been indexed by the search engine and displays links to those pages.

A search for site:hostname.com was performed on each major search engine for each of the sites to be examined, and the number of pages shown as indexed at that search engine noted. If zero or 1 page shown, we designated the site as censored or "banned". For each site we also noted the Alexa rank and Google PR. The page counts reported by Google and Yahoo tend to be higher than the actual number of pages at the larger numbers. For example Google indicates they index 27.6 million pages at the Open Directory, which probably only has about 2 million pages including foreign language pages. (Yahoo indexes 4.2 million, MSN Search 399,000.)

Five of the sites were found to be not operating, moved, no longer using ODP data, or using an unregistered IP number instead of a host name, and were excluded from further examination.

Many of these sites use "live" ODP data obtained from ODP in real time.

Results Summary:

Of the 35 sites in the sample, 17 (49 percent) were censored by at least one search engine. Google banned 13 sites (37 percent), Yahoo banned 4 (11 percent), and MSN banned 3 sites (9 percent). Some additional sites had either an extremely low number of pages indexed or "indexing" confined to "link only" by at least one search engine.

An astonishing 22 (63 percent) of the sites were running Google targeted AdSense ads. The extremely high penetration of Google text ads throughout the web site population gives Google an enormous tracking advantage over the other search engines. See http://www.searchenginehonesty.com/search_engine_spying.html for details.

Analysis

There was no obvious reason concerning site appearance why any of the search engines had chosen to censor any of the particular sites. Banned sites seemed to be of equal or greater quality than uncensored sites. The depth-of-crawl (number of pages indexed) at search engines that did not ban a site did not have any obvious correlation with site traffic or any other observed factor.

The overwhelming impression one gets from looking at this data is that search engine censoring departments are censoring access to sites using ODP data, pretty much without regard to any other consideration (quality of the site, amount of ODP data used, value added, context of use, etc.).

Author's Commentary

It is difficult to believe that ODP is unaware of the search engine issues regarding ODP data. Yet ODP continues to encourage webmasters to use ODP data including the use of software that produces identical clones of the ODP directory. There is no warning in the ODP FAQ to the effect of: "WARNING: Using ODP data will likely result in your entire site being censored by one or more major search engines!" Netscape/AOL is in a somewhat competitive situation relative to the major search engines, which might explain ODP's behavior. Also, each ODP clone has links to ODP on each page and therefore benefits ODP's link popularity, helps with editor recruiting, encourages site submissions, and solicits more webmasters to use ODP data. *Each* ODP clone has some 1.8 *million* links to ODP.

Search engines find it politically impossible to admit that they are censoring sites for using ODP data while using it themselves (Google) and conspicuously not censoring either ODP itself or other major users of ODP data. Therefore there are no warnings regarding ODP data in any of the major search engine's webmaster's guidelines.

The massive censoring by major search engines of sites using ODP data calls into question the entire premise of ODP. Do search engines consider that there is *any* legitimate use of ODP data on a site not owned by them, a partner, or friend? If so, what is that use? If not, ODP's 71,000 volunteer editors are being converted into unpaid, non-stock-optioned, and involuntary employees of Google and other giant companies. How many good editors would volunteer to work for free if they knew they were mainly being used to produce directories for major companies that could easily afford to pay them? The "open" nature of ODP is destroyed if only major search engines and their friends are allowed to use the data without fear of censoring.

It is clear from the data that Google sets their published PageRank to zero for sites that are arbitrarily manually banned by Google. Since Google represents PageRank as

automated, objective, and honest, this appears to be a case of a dishonest, unfair, and potentially litigable (libel and defamation) business practice. See *Google Defamation Case* http://www.searchenginehonesty.com/search_engine_censoring.html for more on this issue.

Search engine censoring is virtually exclusively directed against small businesses. We did not find any case in which a site using ODP data and owned by a larger business was censored by a major search engine.

Related Information

See separate case study of seekon.com, which examines a single censoring case in more depth. This site hosts a small modified subset of ODP data selected to complement other information presented to its intended target audience.

See main site (http://www.searchenginehonesty.com/) for general information on search engines and search engine censoring.

Individual Site Descriptions:

The following list contains the first 40 sites listed on ODP's "Sites using ODP data" page including the title and description by ODP's editor as of March 20, 2006. [The annotations are the number of pages indicated as indexed by each search engine, the Alexa rank, whether site runs Google targeted ads, Google PageRank (PR), and any of our additional notes.]

- AD Chicago - Directory using ODP data restricted to Chicago, Illinois.[Yahoo 26; MSN 486; Google 118; AR >4M; Google ads PR=2]
- ABC.net - Web directory based on ODP RDF data dumps, with rearranged front page. There is also a metasearch tool with options to search the web, the directory, images or shopping. [Yahoo 561,000; MSN 1489; Google 276,000; AR63,000; Google ads PR=5]
- Adilo - Directory using ODP Data. [Yahoo 40,300;MSN banned; Google 31,000; AR1.3M;Google ads PR=2]
- Adilo Directory - Uses ODP directory via DWodp live, with site previews added from Thumbshots. [Yahoo 138,000;MSN 2821; Google banned; AR 1.1M; Google ads; total clone of ODP PR=0]
- Adynasty - Search tool which includes a directory is based on the Open Directory, powered by DWodp, with site previews by Thumbshots. [Yahoo banned; MSN 704; Google 169;AR 3.2M; Google ads; total live clone of ODP PR=2]
- Africatower - African portal with news, weather, currency exchange and directory taken from the Open Directory Regional section for Africa. [Yahoo 83; MSN 166; Google 12,900; AR 835,000; Google ads;large professional site PR=5]
- Alexa - Provides a search powered by Google, and a directory from the ODP. Directory listings can be ordered by popularity, user rating or alphabetically, and there are links to Alexa's information on each site. [Yahoo 322,000; MSN 59,000;

Google 365,000; Alexa does not report rank on their own site!; no Google ads PR=8]

- All About Hosting - Includes a searchable directory of hosting providers, using ODP data. [Yahoo 166; MSN 299; Google 96; AR 3.3M; Google ads; only hosting providers info PR=5]
- Any-hoo - A copy of the Open Directory, powered by DWodp live, with site previews added from Thumbshots, and advertisements from Google. [Yahoo 26,600; MSN 1214; Google 101,000 links only; AR 2.6M; Google ads PR=1]
- Apnaguide - A business directory for India, taken from the business and economy categories for India in the regional section of the Open Directory, powered by DWodp live and with site previews added from Thumbshots. Also offers ODP content in the languages of India. [Yahoo banned; MSN banned; Google 21; AR > 4M; Google ads PR=4]
- 3arab Soft Links Directory - A copy of the Open Directory, powered by phpODP, with site previews by Thumbshots.[Yahoo 16,600; MSN 3,300; Google banned; AR 97,200; Google ads; live ODP clone PR=0]
- ARHS Web Directory - Web directory using ODP Data. [Yahoo banned; MSN banned; Google banned; AR 3.1M; total clone of ODP; Google ads; accepts direct submittal of sites to AHRS PR=0]
- Ariaa - Directory of music sites, using ODP music categories. [Yahoo 187,000; MSN 233; Google 94,700; AR 929,000; Italian; music listings only; no Google ads PR=1]
- Artemotore - Copy of the ODP at http://www.artemotore.com/cgi-bin/pod.pl?dir=/ without proper attribution. [Yahoo 115,000; MSN 2051; Google 452,000; AR 39,000; Live clone of ODP; Google ads PR=5]
- AskComet - Resource for searching the internet. Live data. [Dead link]
- Association of Washington Business: Directory - Washington State's Chamber of Commerce includes a copy of the Open Directory. [Yahoo 9,800; MSN 2.400; Google 178,000; AR 313,000; clone of entire ODP; no Google ads PR=1]
- Best Sites For - Meta search engine using ODP data for directory. [site moved]
- BestSearchers.com - Links to selected search engines, and a directory using live data from the ODP. [Yahoo 131,000; MSN 615; Google 55,100; AR 279,000; Live ODP clone; Google ads PR=4]
- Big Web Directory - Web directory using ODP data. Allows visitors to view in several different color styles. [Yahoo 65,000; MSN 8,600; Google banned; AR 1.8M; total ODP clone; no ads PR=0]
- Bigsearch.ca - Provides a searchable directory based on the Open Directory Project, as well as news from Canadian sources and free space for blogs. [Yahoo 1,470; MSN 5,000; Google 264; AR 986,000; Google ads PR=2]
- BitTimes - Web directory using ODP data, using Phpodp, with site previews by Thumbshots. Heavy advertising added. [Yahoo banned; MSN 651; Google 226; AR 2.7M; Chinese; no Google ads]
- BizMena.com - Searching business resouces of regional category of ODP of Middle East and North Africa with Thumbshots. [Yahoo 15,900; MSN banned; Google 474; Google ads; regional only PR=4]

- Blutensenzen Infozentrum - German health site uses entire ODP directory. [Yahoo 365; MSN 56; Google 1460; AR > 4M; Live ODP clone; no Google ads PR-3]
- Bonus Bot - An ODP-based directory, edited and categorized for shopping, discounts, promotions, travel, city and hotel guides, kids and teens. Includes search by Google. [Yahoo 12,000; MSN 268; Google 690; AR 469,000; Google ads PR=3]
- Business Nation - Uses the ODP data to supplement the business listings in its business library. Live data. Anaconda. [AR 607,000; used unregistered IP address for directory site: http://66.34.238.111/]
- Bytedog - Offers metasearch using Altavista, Alltheweb and Wisenut, automatically filtering bad links, and a directory from the Open Directory. [Yahoo 92,700; MSN 4484; Google banned; AR2M; live clone of total ODP; no ads PR=0]
- Bytehound.com - Uses ODP directory with Thumbshots. [Yahoo banned; MSN 3000; Google Banned; AR 15,000; total ODP clone; Google ads PR=0]
- The CADwire - Delivers CAD industry news. Includes a directory drawn from relevant sections of the ODP. [Yahoo 147,000; MSN 4200; Google banned; AR 173,000; no Google ads; partial ODP data limited to CAD subjects PR=0]
- Canseek.ca - Uses the Regional: North America: Canada category of the ODP data with Thumbshots. [Yahoo 292,000; MSN 2365; Google banned; AR 577,000; Google ads; limited ODP data Canada regional]
- Cape-horn.com - Travel directory for South America and Antarctica drawn from relevant categories in the ODP in English, Spanish and Portuguese. [Yahoo 11.400; MSN 97; Google 46,300; AR > 4M; Live ODP clone regional only; no Google ads PR=2]
- Cat News.com - Cat site uses modified Recreation category of the ODP. [Yahoo 4,000; MSN 256; Google 2,300; AR >4M; Google ads; Live ODP on cat subjects only PR=4]
- CEOExpress - Portal using a webdirectory from the ODP (POD). [No ODP data found]
- Channel Queer - Directory of Irish lesbian and gay sites. Uses live ODP data. [Yahoo 56,800; MSN 1185; Google 34,500; AR <4M; Live ODP; Very small amount of ODP data on site's subject; Google ads PR=4]
- Chillispider - Directory powered by Portal Scripts. [Site not operating]
- Christseek - Uses the complete ODP Directory with Thumbshots, but uses the Christianity subcategory of Society as the starting point. [Yahoo 21,000; MSN 110; Google 23,200; AR >4M; no Google ads; clone of entire ODP PR=1]
- Classifiedv - Directory based on the Open Directory data dumps. [Yahoo banned; AR >4M; MSN 3,000; Google banned; Google ads; clone of entire ODP PR=0]
- Click Position - Contains ODP data and including thumbshots to show previews of listed sites. [Yahoo 164,000; MSN 2,800; Google banned; AR 1.1M; Live ODP clone of entire directory; no ads PR=0]
- CLickheretoleave - Uses entire ODP data. [Yahoo 70,200; MSN 83; Google 70,000; AR 4.1M; clone of entire ODP including adult; Google ads and adult ads PR=0]

- Click2UK Directory - UK search engine which includes the Open Directory data modified by DWodp. [Yahoo 2,900; MSN 1,900; Google banned; AR 683,000; clone of UK regional ODP areas; no Google ads]
- ClixShare.com - Pay per click site uses ODP directory with Thumbshots. [Yahoo 464; MSN 5,800; Google banned; AR 57,700; no Google ads PR=0]

Search Engine Honesty (http://www.searchenginehonesty.com/)