CENTER FOR DEMOCRACY & TECHNOLOGY

Search Privacy Practices: A Work In Progress

CDT Report -- August 2007

Many of the top Internet search companies have recently announced new privacy initiatives aimed at giving users greater control over data about their search activities or stronger assurances that it is being handled appropriately. That the search engines are now competing to provide the best privacy protections is great news for users, who will hopefully see a continuing expansion of choices and controls offered to them for managing the information they share over the Internet.

This report compares the privacy policies that have emerged from the recent spate of announcements. The chart below illustrates how each of the top search engines – AOL, Ask.com, Google, Microsoft, and Yahoo! – deals with the retention of users' search information. The chart is followed by CDT's recommendations for continuing to develop privacy in search and a glossary.

A search engine collects several pieces of information each time a user conducts a search, including the search query itself, the user's Internet Protocol (IP) address, and an identifier stored inside a "cookie" that uniquely identifies a user's Web browser. Search engines may also generate their own information, such as a unique identifier associated with a particular user, Web browser, or computer. Depending on the circumstances, these data elements, alone or in combination with other information, have the potential to identify individual users.

Internet search companies have cited a number of reasons for retaining these kinds of information. Keeping a record of what users search for and click on helps to improve the quality of search results. Search logs are an integral element in delivering sponsored search results and advertising. Retaining search logs also helps to combat fraud and abuse of both the search functionality and associated advertising mechanisms. Each of the search companies tackles all of these issues in its own way, leading each to different conclusions about how much data they need to retain and how long it must be stored.

As search becomes an increasingly essential part of so many Internet users' daily lives, the search engines' recently announced policies begin to place control of sensitive information back into the hands of users, limiting the risk that consumers' personal data will be misused, lost, stolen or otherwise compromised. The chart below illustrates how – and for how long – the top U.S. search companies plan to retain search, IP address, and cookie ID information, and what user controls exist to limit the retention of this data.

SEARCH PRIVACY PRACTICES Companies ordered according to share of U.S. searches.	**How long after search data has been collected will it be removed?**			**How will search data be removed?**			**Is most or all search data shared with a third party on an ongoing basis?**
	IP address	**Cookie ID**	**Query**	**IP address**	**Cookie ID**	**Query**	
Google *Policies will be in place by December 2007, applied retroactively.*	18 months	18 months	Indefinite	Deletes last octet of address.	Deletes partial or complete ID (specifics TBA).	Does not remove.	No.
Yahoo! *Policies will be in place by July 2008. Currently reviewing how to apply policies to historical data.*	13 months	13 months	Indefinite. Some queries will be removed automatically by personal information filter after 13 months.	Deletes last octet(s) of address.	Deletes some portion of ID (specifics TBA).	Applies personal information filter to remove names, SSNs, etc.	No.
Microsoft *Policies will be in place by July 2008, applied retroactively.*	18 months	18 months	Indefinite	Deletes complete address.	Deletes complete ID.	Does not remove.	No.
Ask.com *Policies will be in place in 2007. Currently reviewing how to apply policies to historical data.* For users who opt out of having Ask retain their search data (via AskEraser):	Few hours	Few hours	Few hours	Deletes complete address.	Deletes complete ID.	Deletes complete query.	Shares most query and IP address data with **Google** for provision of sponsored search results. Contractually limits uses of such shared data to providing and improving the partner's specific service and detecting fraud.
For all other users:	18 months	18 months	Indefinite	Deletes complete address or last octet(s) (specifics TBA).	Deletes complete ID.	Does not remove.	
AOL *Policies will be in place in 2007, applied retroactively.*	13 months	13 months	13 months	Deletes complete address.	Deletes complete ID.	Retains only aggregate statistics about search query frequency.	Shares query and IP address data with **Google** for purposes of delivering AOL search and advertising. Contractually limits uses of such shared data to providing and improving AOL's specific service and detecting spam and fraud.

CDT's Recommendations

Towards Increased User Control

One of the recent announcements addressed in the chart is Ask.com's development of its AskEraser product. AskEraser gives users the ability to opt out of having Ask.com retain their search information – including IP address, cookie ID, and search query – beyond a few hours. This gives users who do not want their information stored for many months new control over their searches. Ixquick, a search engine with a vastly smaller share of searches than those in the chart, provides users with a different kind of privacy protection: Ixquick shares search queries with a variety of other search engines and stores them indefinitely itself, but deletes users' IP addresses after 48 hours by default and does not use unique identifiers in its cookies.

On the flip side, all of the companies listed in the chart allow users to create personal accounts by supplying some form of identifying information, such as a name, address, or email address. Google and Ask.com leverage such accounts to provide users with the option of storing their search logs for as long as they want. It is important to note, however, that when users choose to delete information from their personal search history, it will still remain on the search engine's servers until the minimum retention time (18 months for Google and Ask.com) has passed. Thus, this kind of control serves to extend, not limit, the data retained.

Whether or not a search engine offers a personal search history feature, the company still has the ability to correlate a user's account information to his or her search logs. All of the companies listed in the chart currently store account information separately from search information, and some take further steps to limit correlation, but these systems may not be fully privacy protective if re-uniting account information with search information can be easily accomplished.

As these kinds of accounts proliferate and are combined with other services like email, chat, and maps, this question of correlation becomes increasingly important. Giving users true control over which information is linked back to them should be the ultimate goal.

Recommendation: Search companies should continue to work towards providing controls that allow users to not only extend but also limit the information stored about them. As it becomes possible to tie more and more information back to an individual user account, users should control the correlation of their account information with records of their online activities.

Safeguarding Privacy in the Long Term

The chart demonstrates the wide variety of approaches that the search engines take to storing data in the long term. The diversity of techniques used to safeguard information held over long periods reveals that much remains to be learned about how to best address this issue:

- Google removes partial IP address information and partial (or possibly complete) cookie identifiers. Removing this information goes a long way towards reducing the possibility of being able to correlate search queries back to particular users.

- Yahoo! maintains partial IP addresses and partial cookie identifiers, and additionally applies a personal information filter to remove names, addresses, phone numbers, social security numbers, and other personal information that users may have typed in as search terms. In addition, Yahoo! is investigating the use of a non-reversible identifier that is not derived from IP addresses or cookie IDs. Combining this with the application of personal information filters dramatically

reduces the likelihood of being able to correlate search logs to particular users.

- Microsoft takes a different approach, eliminating all unique identifiers. This makes it extremely difficult, if not impossible, to correlate search queries to specific users.

- Ask.com, as previously noted, gives users the option of deleting all of their IP address, cookie ID, and search query information. Information that Ask.com passes on to Google is, however, subject to Google's retention policies.

- AOL retains queries in the aggregate and deletes all IP addresses and cookie identifiers, eliminating the possibility of correlating searches to users. As with Ask.com, information passed to Google is subject to Google's policies.

That each of the top search engines takes its own unique approach to this problem is a positive sign that the companies are actively pursuing ways to better protect privacy. More research is necessary in order to determine which solution or combination of solutions will be most effective in protecting privacy while also serving the business needs of the search companies. As more consumer information moves online, it becomes increasingly important to be able to improve search services without tying searches to particular users and to safeguard the data that must be stored.

Recommendation: **Researchers, academics, and Internet companies should continue to pursue new and innovative methods for (1) improving the quality of search results, preventing fraud and otherwise meeting business needs without tying searches to particular users, and (2) safeguarding data that is stored for long periods.**

The Advertising Balancing Act

As the chart reveals, two of the companies – Ask.com and AOL – rely on a partner to supply search advertising. But for those that supply their own ads, many of their claims about why they need to retain information relate to advertising. Search logs can help the companies understand which ads are most successful for a particular query. More importantly, search logs are necessary in order to measure the performance of ads and bill advertisers. Stored search logs can also be used to investigate fraud and abuses of search advertising systems.

Against the backdrop of these constraints, safeguarding user privacy becomes increasingly important. If search engines *must* retain data for months and months – as many of them claim they need to – storing the data securely, providing notice, giving users choices about how the information is stored, and limiting the retention of the data to specific purposes are essential. Many of the Internet's most amazing innovations are supplied for free thanks to advertising, but the mere presence of advertising-related demands does not justify overlooking privacy concerns. Search engines must balance both.

Recommendation: **Search companies should expand efforts to develop policies that balance the demands of the advertising marketplace with their users' privacy needs. This should include the development of new standards and policies that take privacy into account from the beginning.**

Leveraging Partnerships to Achieve Privacy Aims

The chart highlights the fact that both Ask.com and AOL leverage their contracts with Google to limit Google's uses of their search data. CDT believes that contracts can be extremely powerful in helping to improve privacy across both the search industry and the broader Internet community. If one industry player develops good practices and can require its partners to adopt them as well, it may trigger a wave of positive privacy changes. With respect to search, CDT is hopeful that in the future contractual terms will not only limit uses of data, but retention time as well.

Consumers, when properly empowered through education, can also exert pressure to improve privacy practices. Informed consumers who choose to take advantage of the burgeoning set of tools available to help them control their search information and other online data send a strong message to Internet companies that privacy should be a priority.

Recommendation: **Internet companies should leverage their contracts with partners to promote privacy protections across the board. Consumers can also exert pressure to improve privacy practices by staying informed and making use of available privacy tools.**

Competition Doesn't Replace Need for Meaningful Legislation

The recent privacy announcements by leading search companies represent the best possible form of industry self-regulation, in which companies are actually competing to provide consumers the most robust options for protecting their own privacy. The major search engines have long been competing on the quality of their search results, the clarity of their site design, and their ability to personalize their services. It is high time for privacy to be added to this list.

With no federal law governing how customer information can be used, it has fallen to companies to draft their own privacy guidelines. Unfortunately, industry self-regulation by itself will never provide strong enough privacy safeguards. Some search privacy issues may be addressed, but consumers' personal information will remain vulnerable in many other contexts. In particular, whatever information is retained is available to the government under a mere subpoena, issued without a judge's approval. Companies will continue to face the intricacies and loopholes of our nation's patchwork of privacy laws so long as no federal standard exists. CDT welcomes the rise of true competition in the search privacy space, but no self-regulatory effort can take the place of baseline consumer privacy legislation.

Recommendation: **No amount of self-regulation in the search privacy space can replace the need for a comprehensive federal privacy law to protect consumers from bad actors. With consumers sharing more data than ever before online, the time has come to harmonize our nation's privacy laws into a simple, flexible framework.**

For further information, contact:
Ari Schwartz (202) 637-9800 x107.
Alissa Cooper (202) 637-9800 x110.

Glossary

- *Cookie* – A small text file that a Web site, through the means of a Web browser, saves on a user's computer and retrieves when the user revisits that Web site.

- *Cookie ID* – An identifier stored in a cookie that uniquely identifies a user's Web browser.

- *IP address* – A number assigned to a user's computer by his or her Internet Service Provider (ISP). This number allows Internet communications to be routed to and from the user's computer. Some computers' IP addresses are "static" and do not change over time. Others are "dynamic" and may change as often as every time a user logs on to the Internet.

- *Octet* – A group of eight bits. IP addresses are comprised of 32 bits split into four octets. In an IP address the four octets are usually represented as decimal-separated numbers, each of which must be between 0 and 255 (e.g., 127.0.0.1). The first few octets may reveal some information about where its associated computer is physically located and which ISP assigned the IP address to that computer. There are 256 possible IP addresses using any given set of the first three octets (e.g., the range of IP addresses that begins with "198.6.1" will go from 198.6.1.0 to 198.6.1.255). All of these IP addresses are not necessarily always assigned to particular computers.

- *Query* – The search terms typed in by a visitor to a search site.