Response to the Article 29 Working Party Opinion
On Data Protection Issues Related to Search Engines

Google welcomes the current efforts of the European data protection authorities to address and seek industry perspectives on data protection issues related to search engines, as described in their Opinion WP 148 of 4 April 2008. Google has given a great deal of thought to all of the issues raised by the Working Party in its Opinion. As stated in our letter of 10 June 2007 to the Article 29 Working Party, Google is committed to raising the bar on our own privacy practices for the benefit of Google users. Google is likewise committed to engaging in a constructive dialogue with the Article 29 Working Party and other leading privacy stakeholders around the world. Mr. Turk requested that Google respond to the Working Party's Opinion in writing. This response is an attempt to make a constructive contribution to the discussion on data protection issues related to search engines. In parallel, Google and other companies offering search engine services continue to launch new pro-privacy initiatives and technologies, accelerating a welcome trend towards more competition in good privacy practices. In the interests of transparency, we are following the Working Party's example by publishing this response on our website.

Many of the issues raised in the Opinion are very complex, both in technology and business terms, and have impact outside of data protection matters. To assure an orderly, comprehensive and detailed discussion, we voluntarily opened an in-depth regulatory dialogue with the Irish Data Protection Commissioner, Mr. Billy Hawkes. Since Google's European Headquarters is located in Dublin, and since Dublin is the home of Google's largest group of Europe-based employees, we approached Commissioner Hawkes and his team to explore whether they would be open to working with us as a sort of "lead regulator" on behalf of the Working Party. While we recognize that our work with Commissioner Hawkes does not preclude the competence of any other DPA, we believe it is the most effective and efficient way of conducting in-depth conversations on such a complex topic – something which could not be done well with nearly 30 regulators in 30 different countries in Europe alone, not to mention the rest of the world. Consequently, in late May and early June, eight Google employees, including privacy, policy and technical experts, spent several full days meeting with Commissioner Hawkes and his team to discuss the Opinion with the technical precision those issues deserve. We understand that Commissioner Hawkes circulated his minutes of these meetings to the Working Party.

In making our response herein, Google is proud to announce that we are taking two important steps to improve our privacy practices, consistent with the Working Party's recommendations.

First, the Working Party specifically requested that search engines include a link to their privacy policy on their home page. On July 3, Google added a link to our

Privacy Center on the Google home pages. As we explained in our blog http://googleblog.blogspot.com/2008/07/what-comes-next-in-this-series-13-33-53.html:

> *"So, today we're making a homepage change by adding a link to our privacy overview and policies. Google values our users' privacy first and foremost. Trust is the basis of everything we do, so we want you to be familiar and comfortable with the integrity and care we give your personal data. We added this link both to our homepage and to our results page to make it easier for you to find information about our privacy principles. The new "Privacy" link goes to our Privacy Center, which was revamped earlier this year to be more straightforward and approachable, with videos and a non-legalese overview to make sure you understand in basic terms what Google does, does not, will, and won't, do in regard to your personal information.*
>
> *We think the easy access to our privacy information without any added homepage heft is a clear win for our users and an enhancement to your experience. You can* check out the new Privacy Center here."

We trust that the Working Party will be pleased to see Google has already implemented this change on its home page, and has done so in all its European language-versions.

Second, Google will begin anonymizing the IP addresses in our search logs after 9 months. As you know, Google was the first company in our industry to announce a defined retention period for search logs, and today's announcement marks another significant step in providing Google users with important privacy protections. We will make this change worldwide.

In addition to these advancements in our privacy practices, we respond in greater detail to points in the Working Party's Opinion below. Specifically, we address the concerns outlined in the Opinion regarding: (1) Google's role as a search engine and its compliance with the data protection laws, particularly with regard to the collection and use of IP addresses and cookies; (2) the Legal Framework governing search engines operating on the global Internet platform, including the applicability of Directives 95/46/EC, 2002/58/EC, and 2006/24/EC; and (3) a detailed analysis of lawful purposes for Google's collection and use of logs data, as requested by the Working Party and as we have published on our blog and in the Google Privacy Center.

1. GOOGLE'S SEARCH ENGINE, BUSINESS MODEL & LAWFUL DATA PROCESSING

Google is grateful for the recognition that the Article 29 Working Party gives to the role played by search engines in the development of the information society. Google's mission is to organize the world's information and make it universally accessible and useful, so we are constantly striving to make a significant contribution to the information society. Information is the lifeblood of our 21st century world, and search engines play a significant role in helping people find and use information to improve education, government, economies, and entertainment. At the same time, Google is very conscious of the fact that

our mission must also respect privacy. Without privacy safeguards, the information society cannot succeed.

Google focuses on providing the best user experience possible while respecting user privacy. Individuals can use Google's search services without revealing their real identity (unless they have registered with us, and even such registration can be done pseudonymously). Moreover, Google's search business is offered to the public for free, and is thus inherently superior from a privacy perspective to paid services because it does not require users' real names, billing addresses, credit card numbers or mandatory tax and accounting records. To support this free service, Google primarily relies on being able to serve relevant advertising to its users. The innovative keyword advertising business – which targets ads based on keyword matching and standard log information to determine a user's general location and preferences – uses minimal information to serve effective ads with search results. The availability of free search services to Internet users and the quality of those users' experience depends on being able to legitimately process data to provide the best results to satisfy users' needs.

In its Opinion, the Working Party addresses concerns based on what it identifies as a search engine's two distinct roles: first as a service provider, and second as a content provider. We respond to these concerns below and in greater detail in the Legal Framework analysis in Section 2.

As an initial matter, the Working Party's Opinion expressly refers to search engines' obligation to provide proper notice of the "nature and purpose of their operations," particularly with respect to the type of personal data collected. Google has adopted a policy of personal data minimization so that it only collects, retains and uses information as necessary to provide a safe and robust service in a manner that respects user privacy. Google strives to be as transparent and precise as possible in describing the specific types of information collected in the Google Privacy Policy. The uses made of this information are also set out in detail in the Privacy Policy and other documentation and media publicly available on the Google Privacy Center page. The Google Privacy Center contains an array of information in a manner that is consistent with the Article 29 Working Party Opinion WP 100 of 25 November 2004 endorsing more harmonized information provision, which Google agrees with and supports. The Google Privacy Center includes the following elements:

- A frame with links to the following four key documents: Privacy Overview, Privacy Policy, Privacy FAQ, Privacy Glossary, Privacy Blogs, and Terms of Service.
- A very prominent link to the Privacy Overview and the main Privacy Policy.
- A section with individual links to the pages that describe the specific privacy practices of 29 different products or services.
- A channel on YouTube with numerous privacy videos dealing with subjects like Search Logs, Personalized Search, Street View,

Cookies and many Google services. These videos are meant to explain our privacy practices in simple and comprehensible language to non-technical users, and we're delighted to see that they've been viewed over half a million times already.

At Google we take great care to be transparent about how we use any information we collect from our users – not only in our policies, but in the design of the products themselves – and we continue to experiment with ways to increase this transparency. As such, we welcome any further suggestions that the Article 29 Working Party may wish to make in this regard.

The Opinion also refers to Google as a “content provider” in describing the search engine function of delivering links to relevant web pages *created and published by others* in response to a user's search queries. We respectfully disagree with this characterization of the service and believe it may be helpful to explain how Google's flagship search engine works to clarify our role. Importantly, Google employees do not create, retrieve or group information in response to specific users' queries. Nor does Google seek to "create a new picture" about any topic or any person, as the Opinion alleges. To the contrary, Google's success as a search engine is based on the speed, relevance and comprehensiveness with which it returns search results from the entire World Wide Web. Our search services are necessarily automated, algorithmic and driven by content created and published by others.

Specifically, the software behind our search technology conducts a series of simultaneous calculations requiring only a fraction of a second. It uses more than 200 signals, including our patented PageRank™ algorithm, to examine the entire link structure of the web – currently billions of indexed pages – and determine which pages are most important. PageRank also considers the importance of each page that "casts a vote" or links to a given page, as votes from some pages are considered to have greater value and thus give the linked page greater value. In this sense, our technology uses the collective intelligence of the web to determine a page's importance. Our software then conducts hypertext-matching analysis to determine which pages are relevant to the specific search being conducted. This technology analyzes the full content of a page and factors in fonts, subdivisions and the precise location of each word. It also analyzes the content of neighboring web pages to ensure the results returned are the most relevant to a user's query.

As described above, the Google search engine is not responsible for the creation of content on the web, nor are its search results intended to form a profile of any individual. Rather, Google responds to user search queries with links to what appear to be relevant pages. To the extent an individual is concerned about the content of a given page, his or her complaint is properly directed to the web site hosting that page. Once the webmaster makes these changes, Google's search results will update automatically when it next crawls the page. While Google does

not direct what web pages are cached, much less any personal information that may be contained in those pages, an individual can request to expedite the removal of old webpages after the webmaster makes these changes by contacting Google directly. More information about removing personal information from search results can be found in our Help Center at http://www.google.com/support/bin/topic.py?topic=360.

The Working Party's Opinion also refers to the Common Position on privacy protection and search engines adopted by the International Working Group on Data Protection in Telecommunications on 15 April 1998 and revised on 6-7 April 2006, which indicates that providers of search engines have the capability to draw up a detailed profile of the interests of their users. Google agrees that the impact of individuals' ability to upload personal information for public consumption on the web, as well as the data collected by search engines, deserves proper scrutiny by companies, regulators and the public at large. We also agree that any use of personal information to profile individuals should be done with great care and consistent with established data protection principles. As such, Google strives to operate and provide its services in the most beneficial and least intrusive way for its users.

2. LEGAL FRAMEWORK

Google agrees with and fully supports Article 8 of the European Convention on Human Rights and Fundamental Freedoms and its resonance in the data protection laws in Europe and around the world. Since search engines play a crucial role as a means of accessing information freely, as recognised by the Working Party, it is doubly significant that Google not only takes steps to keep its data secure, but has demonstrated its willingness to challenge overreaching data demands. We make a great effort to reach a sensible balance between our mission to contribute to economic and social progress by offering a remarkable free service to individuals, and our commitment to the respect for human rights and private life in particular. With this commitment in mind, we address the Opinion's guidance regarding the applicability of the Data Protection Directive, Data Retention Directive and ePrivacy Directive below.

2.1 Applicability of Directive 95/46/EC (Data Protection Directive)

The Opinion lays out the Working Party's guidance with respect to the definitions of "personal data" and "controller" as applied to search engines and, in particular companies that offer search engine services on the global Internet platform.

Personal data: IP addresses and cookies

To be very clear, Google has always taken the view that IP addresses should be regarded as confidential information that deserves a very high standard of protection. Indeed, our logs data, which include IP address information, are treated with utmost care and security, accessible to only a small number of engineers who require access for purposes of their job responsibilities. That said, there is significant debate as to whether an IP address should be considered

"personal data" for purposes of data protection obligations. Legal analysis of the potential status of IP addresses as personal data should be as rigorous as possible.

The Working Party refers to its earlier WP 136 opinion regarding Internet service providers and holds that search engines should also treat IP addresses as personal information subject to data protection laws. At the same time, the Working Party acknowledged that in some cases, a service provider does not and cannot identify an individual based solely on an IP address. We agree that the prophylactic guidance of WP 136 must be balanced by a standard of reasonableness." Indeed, Recital 26 of the Data Protection Directive states that determining whether a piece of data is "personal information" requires consideration of all the means likely reasonably to be used..." of identifying the individual. So, sometimes IP addresses should be considered "personal data", for example in the hands of an Internet access provider that attributes those addresses to their own subscriber, whose personal details they hold, like name, address and billing address. On the other hand, IP addresses should not automatically be considered "personal data" in the hands of any website that a user happens to visit, if that website has no ability to identify the user. Google believes that nuanced analysis is required to apply the correct legal characterizations to IP addresses, rather than black-and-white labels.

Google also takes suitable measures to ensure that cookie data remains protected and confidential. Like IP addresses, Google is transparent about its use of cookies and the use of associated data. Users also have the ability to control their cookies through the cookie controls in their browsers. Google does not use Flash cookies for tracking purposes.

Controllers

The Data Protection Directive established the concepts of “controllers” and “processors”, and created specific legal obligations applicable to each. According to the Data Protection Directive, a controller is a natural or legal person, public authority, agency or any other body which alone or jointly with others determines the purposes and means of the processing of personal data. In practice, the key aspect of this definition is the ability to decide how personal data is being collected, stored, used, altered and disclosed. It is possible for this decision to be made jointly by different entities, in which case, they will be regarded as joint controllers. In contrast with the concept of controller, a data processor is a person (other than an employee of the controller) who processes data on behalf of a controller. This distinction is important because the jurisdictional rules that determine the applicability of European data protection law focus on the processing of data carried out by the controller.

Google’s Privacy Policy clarifies that the operational centre where the decisions are made with regard to the processing of user data rests with Google Inc. In particular:

- The Privacy Policy to which all Google users are directed states that the policy applies to all the products, websites and services that are owned and operated by Google Inc. or its subsidiaries or affiliated companies.
- In the Terms of Service referred to in the Privacy Policy, it is made clear that the terms are entered into with Google Inc. and that where any services are provided by affiliates, they are acting on behalf of Google Inc. in providing those services.
- Google states in its Privacy Policy that it adheres to the US Safe Harbor Privacy Principles and that it is registered with the U.S. Department of Commerce's safe harbor program, and the contact address given for queries about the Privacy Policy is in the U.S.
- Google's Privacy Policy also provides that Google processes personal information on its servers in the United States of America and in other countries and that in some cases, personal information is processed on a server not located in the country in which the user is located.

All of these factors indicate that Google Inc. must be regarded as the controller in connection with the processing of users' data irrespective of where the data is collected or stored. Accordingly, Google Inc. – as the parent company of all Google entities – has made a commitment to ensuring that the privacy practices of Google are globally consistent whilst locally compliant.

Article 4 Data Protection Directive / Applicable Law
Taking into account the circumstances where the national data protection laws of each EU member state may be applicable, we would like to set out our understanding of the circumstances where Google may be subject to EU data protection law as a result of either:

- The processing of personal data in the context of an establishment within the EEA; or
- Its operations as a controller of personal data based outside the EEA but using equipment within the EEA.

Establishment on the territory of a Member State (EEA)
Article 4(1)(a) of the Data Protection Directive concerns situations where the data processing operations are carried out in the context of the activities undertaken by the establishment of a controller within the EEA. In the case of Google, if the collection, storage or analysis of search logs or any other associated activity involving the processing of personal data were carried out by one of the Google entities established in the EEA in their capacity as controllers of the information, that entity would be subject to EU data protection law (i.e. the national data protection law of the territory where it is based) in respect of that processing. However, as evidenced above, the fact that a global search engine provider, like Google, has legal entities formed under the law of an EEA member state or branches located within the EEA does not necessarily bring all data processing operations of that search engine provider within the scope of application of EU

law. For that to happen, the EEA-based entity or branch of the search engine provider must (a) be involved in the actual processing of personal data, and (b) do so as a controller.

The Opinion WP 148 refers to several instances where a search engine provider with an establishment within the EEA, would be subject to EU data protection law as a result of the data processing operations carried out in the context of its activities. These include:

- Where an EEA-based establishment is responsible for relations with users of the search engine.
- Where the EEA-based office of a search engine provider is involved in the selling of targeted advertisements to the inhabitants of that state.
- Where the EEA-based establishment of a search engine provider complies with court orders and/or law enforcement requests by the competent authorities of a member state with regard to user data.

However, whilst in these three examples the relevant establishment of the search engine provider may operate within an EEA member state for all sorts of business purposes, the types of activities described may not necessarily involve processing of user data. For example, the practice of selling targeted advertising is unlikely to require any actual processing of user data by the entity arranging the sale. There may also be instances where the local establishment will only process a very limited amount of user data that excludes search-related information – for example, a local EEA-based entity may interface with users of that country in order to deal with enquiries about the service, but that relationship may be restricted to helpdesk-type functions. Finally, there may be cases where all processing of user data by the local EEA-based entity will be carried out in its capacity as a processor for an overseas-based controller, such as Google Inc. Google operates data centres in several locations in the EU (e.g., Belgium, Ireland and The Netherlands). These data centres are used for the purpose of storing data, in particular web index pages, and serving search query results. However, although the EU-based data centres may be managed by a local Google subsidiary, all decisions regarding the purpose and means of the data processing activities are made by Google Inc. In other words, Google Inc. remains the controller of the personal data stored in, and made available from, EU-based data centres. Therefore, as the EU-based Google entities are not acting as controllers for the user data stored in their data centres and it is another entity – such as Google Inc. – that determines the purpose and the means by which the user data is processed, then EU data protection law will not apply to the Google entities established in the countries where the data centres are located in respect of the processing activities taking place in those data centres. In light of the above, to the extent that Google's operations in Europe are purely commercial, providing advice to companies about how to advertise online, or involving research and development, those Google entities that are established in the EEA are unlikely to carry out any processing of Google users' personal

data as controllers of that information. As a result, despite the fact that Google may have establishments within the EEA, given the nature of the commercial activities being undertaken in those establishments, they will not fall within the jurisdiction of EU data protection law as far as the processing of Google users' data is concerned.

Use of equipment
Article 4(1)(c) of the Data Protection Directive concerns situations where the controller of the personal data is outside the EEA but uses equipment to process such data within the EEA. The objective behind this provision is to avoid situations where an individual is not protected as regards processing taking place within his or her country, solely because the controller of the data is not based within the EEA. Given that, as indicated above, Google Inc. will be regarded as the controller of the processing of Google users' data and Google Inc. is not established in the EEA, it is necessary to determine whether in the context of such processing, Google Inc. will be using equipment within the EEA.

One set of circumstances where the Working Party has concluded that foreign search engine providers use equipment within the EEA in this context is where cookie data is stored in, and accessed from, the terminal equipment of an EEA-based user (as identified by the Article 29 Working Party in its opinion WP 56 of 30 May 2002). This is a practice that is acknowledged by Google in its privacy policy where it states that when users visit Google, one or more cookies are sent to their computer in order to identify the browser, store user preferences and track user trends. As the Working Party is aware, this is a nearly universal practice amongst all the world's websites. However, even in this type of case, the wording of the Data Protection Directive is being stretched to cover new technologies that were not envisaged when it was adopted in 1995. Given the nature of cookies and other similar applications that rely on Internet connectivity to enable a service provider to recognise a particular user, concluding that a non-EEA controller is subject to the laws of every EEA member state as a result of the existence of a file in the terminal equipment of its EEA-based users seems very far fetched and beyond the aims of the Data Protection Directive.

As explained above, Google Inc. will have access to any personal data stored in its global network of data centres, including of course those located within the EEA. These cases will certainly involve the use by Google Inc. (as a controller) of equipment situated in the EEA. Therefore, in these cases, the effect of the jurisdictional rule article 4(1)(c) of the Data Protection Directive would likely bring Google Inc. within the scope of application of the national data protection laws of the countries where the data centres are based.
In our view, the effect of the jurisdictional rules that determine the application of European data protection law to the processing of user data by Google is as follows:

- Given the nature of the commercial activities being undertaken by Google's European entities, they will not necessarily fall within the jurisdiction of EU data protection law for global Internet services as far as the processing of users' data is concerned, although they will of course be subject to EU data protection laws regarding their local processing, such as local marketing and HR purposes.
- Despite the argument that foreign search engine providers use the terminal equipment of users within the EEA where cookie data is stored in, and accessed from, such terminal equipment, this Working Party opinion has not been upheld by European courts, and indeed, is not in line with recent European judicial precedents.
- Google Inc. may be subject to the national data protection laws of the EU countries where its data centres are based due to its use of equipment in those countries to store and process user data via local data centres. Notwithstanding these legal, or indeed legalistic, observations, Google is committed to complying with EU data protection principles for the benefit of our users in Europe.

Notwithstanding the fact that the data controller of the processing of Google's users' personal data is Google Inc, as mentioned above, Google's European Headquarters is located in Dublin, and Dublin is the home of Google's largest group of Europe-based employees, we wish to reiterate our willingness to work with Commissioner Hawkes and his team, to reflect the particular nexus between Google in Europe and the country of Ireland.

2.2 Applicability of Directive 2002/58/EC (ePrivacy Directive) and Directive 2006/24/EC (Data Retention Directive)

As mentioned in the Opinion, a search engine – as in the case of Google – may offer additional services (i.e. beyond searches) which may fall within the scope of an "electronic communications service" as defined by the Framework Directive 2002/21/EC. In the context of Google, it remains a subject of much debate whether Google's webmail services such as Gmail would be considered an electronic communications service. We agree with the Working Party that search logs are outside of the scope of the Data Retention Directive.

With regard to the ePrivacy Directive, Google acknowledges the applicability of the relevant articles concerning the use of cookies and the sending of unsolicited communications. Needless to say, Google is entirely confident about its level of compliance with the obligation to provide clear and comprehensive information about the possible use of cookies and the requirements concerning unsolicited communications.

2.3 Google is not a content provider under the Data Protection Directive

Google sincerely welcomes the Working Party's position on freedom of expression and the appropriate balancing with individual right to privacy as set

forth in Article 9 of the Data Protection Directive. However, we also must reiterate the point that Google as a search engine is not a content provider in a legal sense of that term. As mentioned above, Google does not publish or republish content, but provides access to information created and published by others.

The obligations of the Data Protection Directive apply to the processing of personal data carried out by a "controller". A search engine responds to search criteria determined by the individual initiating the search. In analyzing web page content, creating indexes and determining a page's importance in relation to a particular search (as described above), Google is merely providing the means to access information which has been published by others. To the extent that such information contains personal data, those who make the information available will be data controllers. A search engine cannot be said to be determining the purposes for which the personal data was published on the web or the purpose for which the search is initiated. Whilst a search result may contain a set of information about identifiable individuals, Google does not carry out any further processing operation in terms of manipulating the information for presentation in any particular way other than as a list of search results. In other words, Google will be merely operating automatically (although through very sophisticated technology) on the instructions of a user and at most will only provide the means to provide a set of information. Google could be described, therefore, as acting as a processor of information operating on the basis of a set of criteria defined by the person instigating the search. Therefore, Google cannot be regarded as the controller of any personal data included in the search results.

The Opinion argues that the fact that a search engine has "control" of the material is a key factor in determining whether a search engine is a data controller. Control over a database in the sense of being able to remove information on its own initiative or due to a legal requirement does not make the search engine a data controller. The removal of the information would be by way of a block on inclusion of certain information in a search response and the search engine would not have any control over any attempts for the re-publication of the information on a particular website. Google could be required to block or remove information as a result of a legal obligation to which it may be subject under applicable law (e.g. copyright or defamation action or privacy violation), but the ability to do so in respect of information potentially accessible in response to a search does not make Google a controller of that information.

In any event, Google makes available suitable metatags to website owners to allow them to keep the Googlebot from following links to the relevant website. Google obviously will always respect any such robot protocols.
The process of caching is done as part of the indexing process and only involves making a temporary copy of a website until the next time that website is crawled. Google does not make any decisions about the nature of the information being cached and, therefore, is not a controller of any personal data that may exist within that information. As with other issues involving personal information,

this is an area where Google is very willing to work with the Article 29 Working Party to ensure that any privacy concerns in this respect are properly addressed.

3. THE LAWFULNESS OF PROCESSING

Google wishes to emphasise its efforts to explain as clearly and comprehensively as possible the different uses made of any personal data collected. We refer to the information available through the Google Privacy Center mentioned before. In particular, the main Google Privacy Policy describes the purposes for which each type of information collected will be used. The Google Privacy Policy goes on to say that in addition to the above, the purposes will include:

- Providing our products and services to users, including the display of customized content and advertising;
- Auditing, research and analysis in order to maintain, protect and improve our services;
- Ensuring the technical functioning of our network; and
- Developing new services.

This is then expanded on a service by service basis, by referring to the uses made of any personal data collected in the context of the 29 different products or services whose specific privacy practices are described individually.

The different purposes for which Google collects users' information were discussed at length during our meetings with the Irish Data Protection Commissioner and we would like to refer the Article 29 Working Party to the minutes of those discussions.

3.1 Analysis of purposes and grounds by the Working Party

In light of the above, we trust that the level of detail concerning the purposes for which Google may process its users' personal data set out throughout the different components of the Google Privacy Center (including the Google Privacy Policy, the Privacy Overview document, the Privacy FAQ and the pages that describe the specific privacy practices of our different products or services) will meet the expectations of the Article 29 Working Party.

Personal data used by Google in the context of the provision of its services is either expressly provided by our users (e.g., when they create a Gmail account) or necessary for the purposes of Google's legitimate interests (e.g., in the case of any personal data potentially collected from unregistered users). With regard to the former, the legal basis for the processing of the data is clearly the user's consent. With regard to the latter, Google is constantly striving to get the balance right between Google's legitimate interests and the fundamental rights and freedoms of our users. This is obviously consistent with our philosophy to "Be worthy of people's trust" and our respect for users' right to own and control their own data.

In addition, Google has adopted a policy of personal data minimisation so that it only collects, retains and uses information as necessary to provide a safe and robust service in a manner that respects user privacy. We make a virtue of simplicity and this also applies to the information we collect about our users. Data that may pertain to unregistered users may be used to customize a user's experience (for example, using IP addresses to determine the geolocation of a user and deliver a relevant ad from a retailer in the user's area), but we are not interested in identifying unregistered users. Otherwise, we rely on "aggregated non-personal information" (i.e. information that is recorded about users and collected into groups so that it no longer reflects or references an individually identifiable user). For example, this is the case in the context of data shared with advertisers about the aggregate number of users who searched for a particular term or who clicked on a particular advertisement.

Google's pay-per-valid-click business model for advertisers is a significant consideration in justifying data uses. Our model requires the retention of sufficient information to support charges to advertisers. As advertisers are only charged for valid clicks, we need to retain this information for auditing and evidentiary purposes (including cases where we need to provide evidence where a company alleges that it is being charged for fraudulent clicks). Additionally, the retention of logs meets retention standards agreed with auditors to allow for the re-creation of charges to support regulatory (including tax) requirements applicable to Google.

As the Working Party requested, below is more information about the specific uses of logs data. These detailed explanations about our data use were made available to all our users on the Google Blog.

1. System security purposes
Log data is essential to prevent and investigate threats to our users. As we explained in a recent blog post: http://googleblog.blogspot.com/2008/03/using-log-data-to-help-keep-you-safe.html

> *We sometimes get questions on what Google does with server log data, which registers how users are interacting with our services. We take great care in protecting this data, and while we've talked previously about some of the ways it can be useful, something we haven't covered yet are the ways it can help us make Google products safer for our users.*
>
> *While the Internet on the whole is a safe place, and most of us will never fall victim to an attack, there are more than a few threats out there, and we do everything we can to help you stay a step ahead of them. Any information we can gather on how attacks are launched and propagated helps us do so.*
>
> *That's where server log data comes in. We analyze logs for anomalies or other clues that might suggest malware or phishing attacks in our search results, attacks on our products and services, and other threats to our users. And because we have a reasonably significant data sample, with logs stretching back several months, we're able to perform aggregate, long-term analyses that can uncover new security threats, provide greater*

understanding of how previous threats impacted our users, and help us ensure that our threat detection and prevention measures are properly tuned.

We can't share too much detail (we need to be careful not to provide too many clues on what we look for), but we can use historical examples to give you a better idea of how this kind of data can be useful. One good example is the Santy search worm (PDF), which first appeared in late 2004. Santy used combinations of search terms on Google to identify and then infect vulnerable web servers. Once a web server was infected, it became part of a botnet and started searching Google for more vulnerable servers. Spreading in this way, Santy quickly infected thousands and thousands of web servers across the Internet.

As soon as Google recognized the attack, we began developing a series of tools to automatically generate "regular expressions" that could identify potential Santy queries and then block them from accessing Google.com or flag them for further attention. But because regular expressions like these can sometimes snag legitimate user queries too, we designed the tools so they'd test new expressions in our server log databases first, in order to determine how each one would affect actual user queries. If it turned out that a regular expression affected too many legitimate user queries, the tools would automatically adjust the expression, analyze its performance against the log data again, and then repeat the process as many times as necessary.

In this instance, having access to a good sample of log data meant we were able to refine one of our automated security processes, and the result was a more effective resolution of the problem. In other instances, the data has proven useful in minimizing certain security threats, or in preventing others completely. In the end, what this means is that whenever you use Google search, or Google Apps, or any of our other services, your interactions with those products helps us learn more about security threats that could impact your online experience. And the better the data we have, the more effectively we can protect all our users.

2. Fighting Webspam

Data from search logs is also one tool we use to fight webspam and return cleaner and more relevant results. As we explained in a recent blog post: http://googleblog.blogspot.com/2008/06/using-data-to-fight-webspam.html

Webspam, in case you've never heard of it, is the junk you see in search results when websites successfully cheat their way into higher positions in search results or otherwise violate search engine quality guidelines. If you've never seen webspam, here's a good example of what you might see if you click on a link in the search results that's spam (click on the image to see it larger).

You can see how unhelpful such a page would be. This example is filled with almost no original content, irrelevant links, and information that is of little use to a user. We work hard to ensure you rarely see search results like this. Imagine how annoyed you would be if you clicked on a link from a Google search result and ended up on a page like this.

Searchers don't often see blatant, outright spam like this in search results today. But webspam was much more of an issue before Google became popular and before we were able to build effective anti-spam methods. In general, webspam can be a real annoyance, such as when a search on your own name returns links to porn pages as results. But for many searches, where getting relevant information is more critical, spam is a serious problem. For example, a search for prostate cancer that's full of spam instead of relevant links greatly diminishes the value of a search engine as a helpful tool.

Data from search logs is one tool we use to fight webspam and return cleaner and more relevant results. Logs data such as IP address and cookie information make it possible to create and use metrics that measure the different aspects of our search quality (such as index size and coverage, results "freshness," and spam).

Whenever we create a new metric, it's essential to be able to go over our logs data and compute new spam metrics using previous queries or results. We use our search logs to go "back in time" and see how well Google did on queries from months before. When we create a metric that measures a new type of spam more accurately, we not only start tracking our spam success going forward, but we also use logs data to see how we were doing on that type of spam in previous months and years.

The IP and cookie information is important for helping us apply this method only to searches that are from legitimate users as opposed to those that were generated by bots and other false searches. For example, if a bot sends the same queries to Google over and over again, those queries should really be discarded before we measure how much spam our users see. All of this--log data, IP addresses, and cookie information--makes your search results cleaner and more relevant.

If you think webspam is a solved problem, think again. Last year Google faced a rash of webspam on Chinese domains in our index. Some spammers were purchasing large

amounts of cheap .cn domains and stuffing them with misspellings and porn phrases. Savvy users may remember reading a few blogs about it, but most regular users never even noticed. The reason that a typical searcher didn't notice the odd results is that Google identified the .cn spam and responded with a fast-tracked engineering project to counteract that type of spam attack. Without our logs data to help identify the speed and scope of the problem, many more Google users might have been affected by this attack.

In an ideal world, the vast majority of our users wouldn't even need to know that Google has a webspam team. If we do our job well, you may see low-quality results from time to time, but you won't have to face sneaky JavaScript redirects, unwanted porn, gibberish-stuffed pages or other types of webspam. Our logs data helps ensure that Google detects and has a chance to counteract new spam trends before it lowers the quality of your search experience.

3. Preventing Click Fraud

We recently published a blog post to explain better how we use logs data in the fight against click fraud. http://googleblog.blogspot.com/2008/03/using-data-to-help-prevent-fraud.html

Protecting our advertisers against click fraud is a lot like solving a crime: the more clues we have, the better we can determine which clicks to mark as invalid, so advertisers are not charged for them.

As we've mentioned before, our Ad Traffic Quality team built, and is constantly adding to, our three-stage system for detecting invalid clicks. The three stages are: (1) proactive real-time filters, (2) proactive offline analysis, and (3) reactive investigations.

So how do we use logs information for click fraud detection? Our logs are where we get the clues for the detective work. Logs provide us with the repository of data which are used to detect patterns, anomalous behavior, and other signals indicative of click fraud. Millions of users click on AdWords ads every day. Every single one of those clicks -- and the even more numerous impressions associated with them -- is analyzed by our filters (stage 1), which operate in real-time. This stage certainly utilizes our logs data, but it is stages 2 and 3 which rely even more heavily on deeper analysis of the data in our logs. For example, in stage 2, our team pores over the millions of impressions and clicks -- as well as conversions -- over a longer time period. In combing through all this information, our team is looking for unusual behavior in hundreds of different data points.

IP addresses of computers clicking on ads are very useful data points. A simple use of IP addresses is determining the source location for traffic. That is, for a given publisher or advertiser, where are their clicks coming from? Are they all coming from one country or city? Is that normal for an ad of this type? Although we don't use this information to identify individuals, we look at these in aggregate and study patterns. This information is imperfect, but by analyzing a large volume of this data it is very helpful in helping to prevent fraud. For example, examining an IP address usually tells us which ISP that person is using. It is easy for people on most home Internet connections to get a new IP address by simply rebooting their DSL or cable modem. However, that new IP address will still be registered to their ISP, so additional ad clicks from that machine will still have something in common. Seeing an abnormally high number of clicks on a single publisher from the same ISP isn't necessarily proof of fraud, but it does look suspicious and raises a flag for us to investigate. Other information contained in our logs, such as the browser type and operating system of machines associated with ad clicks, are analyzed in similar ways.

These data points are just a few examples of hundreds of different factors we take into account in click fraud detection. Without this information, and enough of it to identify fraud attempted over a longer time period, it would be extremely difficult to detect invalid clicks with a high degree of confidence, and proactively create filters that help optimize advertiser ROI. Of course, we don't need this information forever; last year we started anonymizing server logs after 18 months. As always, our goal is to balance the utility of this information (as we try to improve Google's services for you) with the best privacy practices for our users.

4. *Improving Search Quality*

We recently published a blog post explaining how we use search logs to improve search quality: http://googleblog.blogspot.com/2008/03/making-search-better-in-catalonia.html

One of the most important uses of data at Google is building language models. By analyzing how people use language, we build models that enable us to interpret searches better, offer spelling corrections, understand when alternative forms of words are needed, offer language translation, and even suggest when searching in another language is appropriate.

One place we use these models is to find alternatives for words used in searches. For example, for both English and French users, "GM" often means the company "General Motors," but our language model understands that in French searches like seconde GM, it means "Guerre Mondiale" (World War), whereas in STI GM it means "Génie Mécanique" (Mechanical Engineering). Another meaning in English is "genetically modified," which our language model understands in GM corn. We've learned this based on the documents we've seen on the web and by observing that users will use both "genetically modified" and "GM" in the same set of searches.

We use similar techniques in all languages. For example, if a Catalan user searches for resultat elecció barris BCN (searching for the result of a neighborhood election in Barcelona), Google will also find pages that use the words "resultats" or "eleccions" or that talk about "Barcelona" instead of "BCN." And our language models also tell us that the Estonian user looking for Tartu juuksur, a barber in Tartu, might also be interested in a "juuksurisalong," or "barber shop."

In the past, language models were built from dictionaries by hand. But such systems are incomplete and don't reflect how people actually use language. Because our language models are based on users' interactions with Google, they are more precise and comprehensive – for example, they incorporate names, idioms, colloquial usage, and newly coined words not often found in dictionaries.

When building our models, we use billions of web documents and as much historical search data as we can, in order to have the most comprehensive understanding of language possible. We analyze how our users searched and how they revised their searches. By looking across the aggregated searches of many users, we can infer the relationships of words to each other.

Queries are not made in isolation – analyzing a single search in the context of the searches before and after it helps us understand a searcher's intent and make inferences. Also, by analyzing how users modify their searches, we've learned related words, variant grammatical forms, spelling corrections, and the concepts behind users' information needs. (We're able to make these connections between searches using cookie IDs –

small pieces of data stored in visitors' browsers that allow us to distinguish different users. To understand how cookies work, watch this video.)

To provide more relevant search results, Google is constantly developing new techniques for language modeling and building better models. One element in building better language models is using more data collected over longer periods of time. In languages with many documents and users, such as English, our language models allow us to improve results deep into the "long tail" of searches, learning about rare usages. However, for languages with fewer users and fewer documents on the web, building language models can be a challenge. For those languages we need to work with longer periods of data to build our models. For example, it takes more than a year of searches in Catalan to provide a comparable amount of data as a single day of searching in English; for Estonian, more than two and a half years worth of searching is needed to match a day of English. Having longer periods of data enables us to improve search for these less commonly used languages.

At Google, we want to ensure that we can help users everywhere find the things they're looking for; providing accurate, relevant results for searches in all languages worldwide is core to Google's mission. Building extensive models of historical usage in every language we can, especially when there are few users, is an essential piece of making search work for everyone, everywhere.

5. Using Logs to Improve the Science of Search

We recently published a blog post explaining our philosophy that better data makes for better science: http://googleblog.blogspot.com/2008/03/why-data-matters.html

Better data makes for better science. The history of information retrieval illustrates this principle well. Work in this area began in the early days of computing, with simple document retrieval based on matching queries with words and phrases in text files. Driven by the availability of new data sources, algorithms evolved and became more sophisticated. The arrival of the web presented new challenges for search, and now it is common to use information from web links and many other indicators as signals of relevance.

Today's web search algorithms are trained to a large degree by the "wisdom of the crowds" drawn from the logs of billions of previous search queries. This brief overview of the history of search illustrates why using data is integral to making Google web search valuable to our users.

A brief history of search
Nowadays search is a hot topic, especially with the widespread use of the web, but the history of document search dates back to the 1950s. Search engines existed in those ancient times, but their primary use was to search a static collection of documents. In the early 60s, the research community gathered new data by digitizing abstracts of articles, enabling rapid progress in the field in the 60s and 70s. But by the late 80s, progress in this area had slowed down considerably.

In order to stimulate research in information retrieval, the National Institute of Standards and Technology (NIST) launched the Text Retrieval Conference (TREC) in 1992. TREC introduced new data in the form of full-text documents and used human judges to classify whether or not particular documents were relevant to a set of queries. They released a sample of this data to researchers, who used it to train and improve their systems to find

the documents relevant to a new set of queries and compare their results to TREC's human judgments and other researchers' algorithms.

The TREC data revitalized research on information retrieval. Having a standard, widely available, and carefully constructed set of data laid the groundwork for further innovation in this field. The yearly TREC conference fostered collaboration, innovation, and a measured dose of competition (and bragging rights) that led to better information retrieval.

New ideas spread rapidly, and the algorithms improved. But with each new improvement, it became harder and harder to improve on last year's techniques, and progress eventually slowed down again.

And then came the web. In its beginning stages, researchers used industry-standard algorithms based on the TREC research to find documents on the web. But the need for better search was apparent--now not just for researchers, but also for everyday users---and the web gave us lots of new data in the form of links that offered the possibility of new advances.

There were developments on two fronts. On the commercial side, a few companies started offering web search engines, but no one was quite sure what business models would work.

On the academic side, the National Science Foundation started a "Digital Library Project" which made grants to several universities. Two Stanford grad students in computer science named Larry Page and Sergey Brin worked on this project. Their insight was to recognize that existing search algorithms could be dramatically improved by using the special linking structure of web documents. Thus PageRank was born.

How Google uses data

PageRank offered a significant improvement on existing algorithms by ranking the relevance of a web page not by keywords alone but also by the quality and quantity of the sites that linked to it. If I have six links pointing to me from sites such as the Wall Street Journal, New York Times, and the House of Representatives, that carries more weight than 20 links from my old college buddies who happen to have web pages.

Larry and Sergey initially tried to license their algorithm to some of the newly formed web search engines, but none were interested. Since they couldn't sell their algorithm, they decided to start a search engine themselves. The rest of the story is well-known.

Over the years, Google has continued to invest in making search better. Our information retrieval experts have added more than 200 additional signals to the algorithms that determine the relevance of websites to a user's query.

So where did those other 200 signals come from? What's the next stage of search, and what do we need to do to find even more relevant information online?

We're constantly experimenting with our algorithm, tuning and tweaking on a weekly basis to come up with more relevant and useful results for our users. But in order to come up with new ranking techniques and evaluate if users find them useful, we have to store and analyze search logs. (Watch our videos to see exactly what data we store in our logs.) What results do people click on? How does their behavior change when we change aspects of our algorithm? Using data in the logs, we can compare how well we're doing now at finding useful information for you to how we did a year ago. If we don't keep a history, we have no good way to evaluate our progress and make improvements.

To choose a simple example: the Google spell checker is based on our analysis of user searches compiled from our logs – not a dictionary. Similarly, we've had a lot of success in using query data to improve our information about geographic locations, enabling us to provide better local search.

Storing and analyzing logs of user searches is how Google's algorithm learns to give you more useful results. Just as data availability has driven progress of search in the past, the data in our search logs will certainly be a critical component of future breakthroughs.

Conclusion

At Google, we know that respecting our users' privacy is fundamental to earning and keeping their trust. And limiting the amount of time that our raw logs are retained is an important privacy protection for our users. Last year, we announced (http://googleblog.blogspot.com/2007/06/how-long-should-google-remember.html) that we would anonymize our logs after 18 months. At the time, we were the first in our industry to announce such a practice. We have now decided to go farther and begin anonymizing the IP addresses in our logs after just nine months. We haven't sorted out all of the implementation details, but we have committed to making it work. We are of course committed to continuing an on-going discussion with the Working Party, as we are able to refine the technical implementation details.

Finding the right balance between data retention and privacy is a tough issue for policymakers, Google and our industry. There is great utility in data, but we also believe that limiting the amount and types of data we keep can improve privacy while continuing to provide a strong user experience. Anonymizing the data earlier will have costs, particularly in terms of future search quality improvements. But our engineers are working hard to minimize those losses.

We trust that the Working Party will welcome these significant steps. We thank the Working Party for its on-going leadership and contributions to the important public, worldwide debate about privacy and the Internet.

Peter Fleischer
Global Privacy Counsel
Google