ALBERT GIDARI, JR., *pro hac vice*
(AGidari@perkinscoie.com)
PERKINS COIE LLP
1201 Third Avenue, Suite 4800
Seattle, WA 98101
Telephone: (206) 359-8000
Facsimile: (206) 359-9000

LISA A. DELEHUNT, Bar No. 228551
(LDelehunt@perkinscoie.com)
PERKINS COIE LLP
180 Townsend Street, 3rd Floor
San Francisco, California 94107-1909
Telephone: (415) 344-7000
Facsimile: (415) 344-7050

Attorneys for Respondent
GOOGLE INC.

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

SAN JOSE DIVISION

ALBERTO R. GONZALEZ, in his official capacity as ATTORNEY GENERAL OF THE UNITED STATES, Movant, v. GOOGLE INC., Respondent.	CASE NO. 5:06-mc-80006-JW DECLARATION OF MATT CUTTS

MATT CUTTS states as follows:

I. INTRODUCTION

1. I am a Senior Staff Software Engineer and head of the Webspam group at Google. I have held this position since November 2004. I have worked with Google since February 2000 as a technical engineer in one capacity or another. I have personal knowledge of the facts set forth below and, if called as a witness about those facts, could testify competently thereto.

2. Currently, I manage the Webspam group. "Webspam" does not refer to spam emails, but rather to pages from the World Wide Web ("Web") that violate Google's quality guidelines in an attempt to rank higher on Google. For example, a page containing pornographic material that attempted to show up as a result to a search for "Disney cartoons" would be considered webspam. Accordingly, I am knowledgeable about how Google crawls, indexes, and ranks web pages. I have written a variety of programs for Google, including programs having to do with filtering pornography and detecting web pages that attempt to bypass Google's quality guidelines. I have a Master of Science in Computer Science, *summa cum laude*, from University of North Carolina at Chapel Hill. I am recipient of a National Science Foundation Graduate Research Fellowship, a Link Foundation Fellowship, and a Gaines Fellowship in the Humanities. I also have a Bachelor of Science in Computer Science and Mathematics, *summa cum laude*, from the University of Kentucky. I have co-authored and published a variety of articles relating to computer graphics and computer vision.

3. I have reviewed the Motion to Compel Compliance with Subpoena Duces Tecum brought by Alberto R. Gonzales (the "Government") against Google, Inc. and the supporting declarations and exhibits, including the Declaration of Professor Philip B. Stark, Ph.D. The Government has requested that Google produce a multistage random sample of one million URLs from Google's databases. URL stands for "Uniform Resource Locator" and acts as the global address of documents and other resources on the World Wide Web ("Web"). URLs consist of the protocol (e.g. HTTP) and the Internet Protocol ("IP") address or domain name. The Government has suggested that Google select at random 100 of its data centers containing URLs, and then

select at random 10,000 URLs from each of those data centers. In his declaration, Professor Stark has stated that he could be involved in the random selection process.

4. Additionally, the Government has requested that Google produce every single search query entered at www.google.com over a given week.

II. SUMMARY OF DECLARATION

5. In my declaration, I testify as to the following:

A. In Section III, I summarize how Google's search engine works.

B. In Section IV, I explain how examining the search queries Google's users have typed into Google's search engine at www.google.com does not reveal accurate information about the likelihood of minors' exposure to material the Government deems Harmful to Minors ("HTM") or about the search patterns of users.

C. In Section V, I explain that there is little to no correlation between a URL's presence in Google's index and if or when that URL would be returned as a result responsive to a specific set of query words.

D. In Section VI, I discuss the fact that, in order to protect the privacy of its users, Google does not disclose to third parties the raw records of searches entered into its search engine.

E. In Section VII, I explain that Google does not disclose the queries it receives from its users because it offers information about the nature of Google's users and of Google's market share in both the United States and other countries and languages, which could allow Google's competitors to compete more effectively with Google. Additionally, Google does not disclose its index of URLs or any such large samples from its index, because one could glean from a sample of the URLs from Google's index confidential and proprietary information about how Google's search engine works.

F. In Section VIII, I explain those measures of which I am personally aware that Google undertakes to protect the confidentiality of its proprietary information.

G. In Section IX, I discuss reasonable, viable alternative public sources of the information the Government is seeking from Google.

H. In Section X, I explain the burden that would be imposed on Google if it were required to respond to the Government's subpoena.

III. GOOGLE'S SEARCH ENGINE

6. Google provides the world's most-used search engine at www.google.com. In general terms, search engines allow users of the Web to search a considerable portion of the vast amount of material located on the Web by entering text queries into the engine. The Web is composed of billions of publicly accessible web sites from around the world and other information sources that web browsers can access. Search engines offer this search capability by creating an index of certain Web content through the use of programs that "crawl" the Web and automatically fetch Web pages. The search engine then allows users to search that index of certain Web content through an interface such as that found at www.google.com. When a user enters a query, the engine searches its Web index and retrieves results relevant to the query. Google's search engine is open to the public. Google is the leader among the major search engines (including Yahoo!, Ask Jeeves, and MSN Search) to a large degree because of its sophisticated and proprietary technology that returns the most relevant and useful results in response to user queries.

7. Some of Google's services, like Google News, allow users to store or establish repeat search queries. Google users may set up recurring searches with results sent to email accounts at defined intervals. Google then processes these requests for its users and sends the results to the email accounts, wireless phones, or other designated mobile devices.

IV. SEARCH QUERIES

8. Examining the search queries Google's users have typed into Google's search engine at www.google.com does not reveal accurate information about how likely it is that minors might be exposed to material the Government deems Harmful to Minors ("HTM") or about the search patterns of users.

9. First, queries entered into Google's search engine have no easily computed correlation to how likely it is that any user might be exposed to what the Government might designate as HTM. This is because Google uses a number of proprietary and confidential methods to return search results to users. Over a period of years, and at enormous expense, Google has developed and implemented multiple unparalleled technologies to crawl and index a considerable portion of the Web and rank the results so as to ensure that Google's users receive the most relevant search results possible. Google employs proprietary methods to determine which documents to crawl from the web and sophisticated techniques to decide which documents to retrieve for a query. Google uses a combination of algorithms and a unique network architecture (among other functionalities) to accomplish this. Most of these crawling, collecting, and sorting techniques are not known outside Google, and they are particularly kept secret from competitors.

10. Additionally, users entering the same query at different computers may receive different search results. The search results presented to the user may vary based on whether SafeSearch (Google's pornography filter) is enabled for that user, the query language requested by the user, or whether the user belongs to Google's program to deliver personalized search results. Google employs additional proprietary techniques that can return different results to different users for the same query. In short, a query alone is not sufficient to determine the results that Google will present to a user.

11. There is no way to tell what queries have been entered by minors and what queries have been entered by adults.

12. Moreover, the Government's request for search queries in an attempt to understand how individuals use search engines is flawed because it does not distinguish between "real" queries entered by individual users, automatic queries generated by a program, known as "bot" queries, and artificial queries generated by real users.

13. For example, a single person may write a short program to send the query "underage sex resources" hundreds, thousands, or even tens of thousands of times a day to search engines using a program, or "bot query." I am also aware of individuals who have written

programs that detect whether a computer is connected to the Web by sending periodic queries to Google; such programs can send thousands of queries to Google each day. Google also receives a fraction of its queries from malicious programs, such as the Santy worm.

14. In addition to queries sent by "bots," test programs, and worms, an individual web page owner (also known as a "Webmaster") may send dozens of queries by hand to check on how his or her websites rank in Google. Many website owners check their rankings every day, which can cause further skew in the query log. I am also aware of efforts by some Google users to deliberately send pornography queries to Google in reaction to the Government's subpoena. One individual wrote a feature for the popular Firefox web browser that will send a random pornography query to Google whenever a user performs a normal query, as if the pornography query had also been entered by the user. Attached as Exhibit A is a true and correct copy of this blog entry which can be found at http://www.hughes-family.org/wordpress/2006/01/23/help-the-justice-department-out-via-greasemonkey/, and which was printed on February 16, 2006.

15. Without removing these artificial and automatic queries, one week of raw query logs will be skewed beyond usability for many purposes. It would be difficult—perhaps impossible—for most researchers to distinguish between such artificial queries and real queries without considerable effort and the use of proprietary techniques.

16. Google's proprietary techniques for returning search results are not static, and Google's algorithms change regularly. Thus, the identical search query run in Google's search engine today is likely to yield different search results than an identical search conducted yesterday, last week, or last month.

V. GOOGLE'S INDEX OF URLS

17. The presence of a URL in Google's index is not in any way representative of the frequency that the URL will be shown to a user. There is little to no correlation between a URL's presence in the index Google maintains of copies of documents it has collected from crawling the web and if or when that URL would be returned as a result responsive to a specific set of query words. Using its proprietary and confidential technology, Google employs more than 100 factors

in scoring documents for relevancy besides just the document's URL. Some of the factors are straightforward. For example, Google considers whether query words are present in the title of the document, how many times the query words appear in the document, the proximity of the query words to each other, as well as a document's PageRank, which is Google's patented method of measuring the reputation of a page. Most of Google's scoring factors and how they are combined is confidential.

18. Beyond scoring factors, some documents in Google's index will have additional demotions or will be blocked for webspam or legal reasons. For example, Google's index contains documents for which we have received valid complaints under the Digital Millennium Copyright Act, and Google does not return those documents in our search results.

19. For all these reasons, it does not follow that a sample of URLs from Google's index will indicate how often searches will return HTM material. Simply put, the fact that a URL in Google's index may appear to contain content the Government considers HTM is not representative of whether and how often HTM content appears as a result of a query. For example, if 5 percent of the URLs listed in Google's index contain HTM content, that fact alone does not mean that any given search would yield 5 percent HTM content.

20. The adage "you can't judge a book by its cover" applies on the Web: URLs alone do not indicate whether content at that URL is HTM. A URL may or may not be logically connected to its content. Sites may have names that suggest explicit sexual material, but actually do not contain HTM content. A URL such as http://www.pbs.org/wgbh/pages/prontline/shows/porn/etc/links.html contains the word "porn" but provides links to anti-pornography organizations such as the American Family Association and the Family Research Council. Likewise, a URL such as http://www.porn-free.org/ may appear pornographic from the URL, but is a faith-based site that is against pornography. The reverse is also true: URLs may have name that seem innocent but actually contain material the Government might consider HTM. The classic example was that www.whitehouse.gov was a harmless website about the White House, while, until recently, www.whitehouse.com contained pornographic

material. Furthermore, the content of web pages at URLs is fluid and dynamic, and the actual content can be difficult to ascertain. For example, until recently the URL http://www.crisiscentersyr.org was evidently owned by a rape crisis center. It appears that the domain expired in late 2005 and was instead registered by someone else. The new owner kept the appearance of the site, making it appear it still belongs to a rape crisis center, but added links to pornography at the bottom of the document. In September 2005, that URL contained non-pornographic material, while in January 2006, it contained links to pornographic material. In short, a URL alone is insufficient to determine whether a document is harmful to minors. This example also shows the dynamic nature of Web content. The content of a web page can change at anytime, and some Web pages update their content daily or more frequently.

21. Additionally, in an attempt to rank higher in Google's search results, some Webmasters show different results to Google than they do to individual users. That is, Google uses a program to crawl the Web to find Web pages for its index. Google's program comes from a specific set of IP addresses and identifies itself as "Googlebot." Some Webmasters program their Web pages to show certain content to Google, content that may be innocent and may seem relevant to a search query. The Webmaster, however, shows different content to individual users, and the content may be pornographic or material otherwise prohibited by Google's guidelines. Such "bait and switch" tactics are called "cloaking." Ultimately, through its diligence, Google finds and removes those documents, but the nature of these Web pages is another example of how a URL does not indicate what content will be displayed in response to a search query.

22. The presence of URLs in Google's index is not reflective of the entire Web. Google uses its technology to crawl and index only the best documents from the Web. The Web can be viewed as having an infinite number of pages. For example, a single web server running a calendar application could generate dynamic web pages with dates going forward forever. Therefore, Google's mission must be to retrieve and index the most useful pages that it can from the infinite number of potential pages that can be retrieved.

23. The only way to get a current and accurate snapshot of the search results that would be returned by a query into Google's search engine would be to run the query on Google's search engine. Notably, running millions of queries on Google above and beyond the normal use of Google would burden Google's system and possibly shut it down. Alternatively, one would need to understand how Google crawls the web, collects URLs, sorts them, indexes them, ranks them, and returns them as search results—in other words, exactly how Google's crown-jewel trade secrets function.

VI. SEARCH QUERIES AND PERSONALLY IDENTIFYING INFORMATION

24. Google does not publicly disclose the searches queries entered into its search engine. If users believe that the text of their search queries into Google's search engine could become public knowledge, they may be less likely to use the search engine for fear of disclosure of their sensitive or private searches for information or websites.

25. There are ways in which a search query alone may reveal personally identifying information. For example, many internet users have experienced the mistake of trying to copy-and-paste text into the search query box, only to find that they have pasted something that they did not intend. Because Google allows very long queries, it is possible that a user may paste a fragment of an email or a document that would tie the query to a specific person. Users could also enter information such as a credit card, a social security number, an unlisted phone number or some other information that can only be tied to one person. Some people search for their credit card or social security number deliberately in order to check for identity theft or to see if any of their personal information is findable on the Web.

VII. GOOGLE'S COMPETITIVE AND CONFIDENTIAL INFORMATION

26. Another reason that Google does not disclose the queries it receives from its users is because it offers information about the nature and effectiveness of Google's search engine and Google's market share in the United States and other countries, which could allow Google's competitors to compete more effectively with Google. Disclosing an entire week's worth of queries would give an estimate of the number of queries that Google processes, which could be

used to deduce market share among search engines. Google protects this information not only by not disclosing search queries but by not disclosing even the amount of computers Google maintains to run its search engine. A week's worth of queries would also indicate the percentage of queries Google tends to receive in each language, which would allow competitors to estimate Google's relative market share in a given language or country, to learn opportunities for market growth, and to decide where to allocate resources for each language or country. The data would also indicate the average length of query that Google's search engine typically receives and how Google's users search, such as what percentage of queries use special search operators or punctuation. This information could allow competitors to better understand the type of users who seek out Google or to allocate resources on specific search operators. Similarly, the queries could relate information regarding whether users tend to use Google for navigational help, for research, or for shopping.

27. Google also does not disclose its index or such a large sample from its index. One could estimate from a sample of the URLs from Google's index information such as (i) the size of Google's index; (ii) how "deeply" Google crawls in different countries or languages (i.e., how many URLs are crawled from each website on average); and (iii) the ability of Google's crawl metrics to measure the reputation of pages or domains. For example, the amount of crawling between different top-level domains such as .com and .uk compared to .pl and .jp would disclose much about how Google's crawling works. The depth of the crawl, the languages of the URLs crawled, and the number of distinct sites crawled could all reveal confidential information about Google's technology. As described above, Google uses multiple methods for crawling the Web, collecting URLs, indexing them, ranking them, and providing relevant results to Google users. Google developed these methods over a number of years and at considerable expense. As far as I know, these methods are not known by Google's competitors. These methods are critical to Google's success as the world's leading search engine and popularity as the world's most-used search engine.

28. While Professor Stark has not disclosed what he intends to do with Google's data or how he intends to do it, it seems logical that Professor Stark will need personal knowledge of how the URLs are collected and maintained. Given my experience in this area, I believe that to obtain that level of knowledge, Professor Stark will need to understand how Google crawls the Web and indexes information. In addition, anyone wishing to attack the randomness of the sample, Professor Stark's methodology (which he has not disclosed), or his ultimate opinions might want to probe into everything that goes into maintaining Google's URL database. For example, it might be important to defend or attack Professor Stark's methodology to know what percentage of queries are initiated within the United States, what percentage of the Web is found on Google's URL index, and how a URL is returned as a result to a search query.

VIII. GOOGLE'S MEASURES TO MAINTAIN THE CONFIDENTIALITY OF ITS INFORMATION

29. I am personally aware of some of the measures Google takes to maintain the confidentiality of its query log, index, and proprietary crawling, indexing, and retrieval technology. The methods of which I am aware are described below. Additional methods are described in the Declaration of Marty Lev.

30. The Google buildings that I have visited all have secured doors and entrances; I have no reason to think that any Google building that houses the information I describe above would be unsecured. Employees and visitors are required to carry appropriate security badges that security personnel routinely check. Certain entrances are restricted by access devices. There are also security guards on duty in the building.

31. Google's information is compartmentalized so that only employees with a need to know have access to certain information. Thus, the information that a sales employee can access is different from the information an engineer can access. For example, sales employees do not have access to Google's source code or to certain parts of Google's technical information. Access to Google's computers is controlled by employee logins and passwords, and employee logins allow them to access only the information they need and blocks them from sensitive information

they do not need. Interns, independent contractors, and temporary employees require special clearance before they may access parts of Google's technical information.

32. Access to Google's index is restricted to engineers. I believe that most employees do not even know the specifics of where the index is stored. The index cannot be viewed in decipherable form unless one uses a source code compiled with specific Google libraries. Additionally, Google is implementing a system where its most sensitive data—such as documentation reflecting the methods it uses to measure the reputation of web pages—will be taken off line and kept in hard copy in a locked location, and an employee wanting access will have to request such access and sign out the file.

33. Likewise, access to Google's query log is particularly restricted to a much smaller group of employees with special clearance who need access to perform their job duties. These employees must reaffirm their need for access to the query log periodically, and anyone who does not reaffirm that need loses their access automatically.

34. If a Google employee does not have the correct security clearance for the query log, he or she must file a request to get it, explaining in writing his or her need to know and promising to keep the information confidential. I am aware that any abuse of Google's internal security and privacy policies is grounds for immediate termination.

35. I am aware that Google has refused requests for search queries in the past. For example, a professor by the name of Amanda Spink at the University of Pittsburgh, requested access to query data from Google and her request was denied. Additionally, Google generally does not share its index.

IX. ALTERNATIVE SOURCES

36. Both query data and web documents are available from several other sources. Metasearch engine Dogpile provides a service called "SearchSpy" to see queries done on that engine at http://www.dogpile.com/info.dogpl/searchspy/. Metasearch engine MetaCrawler provides a service called "Metaspy" which provides a similar list of queries at http://www.metacrawler.com/info.metac/searchspy. The search engines Infospace and

WebCrawler also provide similar information. For under $250 per year, Wordtracker.com sells a database of over 330 million queries which is updated on a weekly basis. Ask Jeeves recently offered a service called "Ask Jeeves Take a Peek" at http://www.ask.com/docs/peek/ which showed questions being typed at Ask Jeeves, and the page would refresh automatically twice a minute. According to her website, Prof. Spink successfully obtained query data from Excite, AltaVista, Ask Jeeves, Fast/AllTheWeb, Vivisimo, Dogpile, Metacrawler, Webcrawler, and Infospace. Prof. Spink has also written a book discussing trends and characteristics of user interaction with search engines.

37. An index of web documents is also available from other sources. One approach would be to crawl the public web, and several research groups have done that. In addition, an Amazon.com subsidiary, Alexa.com ("Alexa"), recently released a service that allows anyone to access Alexa crawl data. The system allows users to process over 4 billion URLs and over 1.7 billion full-text documents. The system is specifically intended for "[r]esearchers who wish to tackle problems related to Web content," according to http://pages.alexa.com/awsp/docs/WebHelp/Introduction/Who_Should_Use_the_Platform.htm. Alexa.com also provides tools to search, process, and publish one's own custom subset of data. A researcher or developer could use this system to test code, including pornography filters, over the full text of documents (not just URLs) and to test code with much more than one million URLs. Attached as Exhibit B is a copy of a web log entry about Alexa which can be found at http://battellemedia.com/archives/002116.php, and which was printed from the internet on February 16, 2006.

38. I believe that Alexa offers Professor Stark a reasonable and viable option to accomplish his goals as set forth in his declaration. I believe that Alexa data would allow for more thorough testing than a sampled list of URLs from Google, as, according to Alexa, their system provides over 300 terabytes of data for researching and testing.

X. THE DEMAND ON GOOGLE TO COMPILE THE INFORMATION REQUESTED

39. Google does not index URLs in the form the Government has requested. Rather, Google maintains an index of copies of the billions of web pages that it has crawled. The index requires vast amounts of computer space and is maintained on multiple computers. There is no existing method to simply copy and hand over in electronic or paper form a selection of URLs from Google's index. The type of information sought by the Government is not created or used internally.

40. Accordingly, an engineer would need to write a computer program capable of gathering the URLs that the Government seeks. The engineer would need to develop the code by determining a scheme to map the data logs and pull random URLs from it. The code would then need to be implemented, which involves a potentially lengthy debugging process in which the engineer tests portions of the code (and eventually the entire program) to identify and fix problems. It is common for debugging to take longer than the initial coding. Finally, the selected documents must be extracted somehow and copied in a form that can be provided to the Government. I estimate that this will take two to five full-time days of an engineer's time. This procedure could be further complicated if Professor Stark needs to participate to verify the randomness of the sample. Google does not dedicate any engineers for this type of task, and so an engineer would have to be diverted from his or her normal job responsibilities.

41. Similarly, Google also does not maintain search queries in the form requested by the Government. The query logs maintained by Google will have to be scrubbed of any personally identifying information. Thus, as in searching the URLs, responding to this request will require writing and implementing new code. I estimate that this will take one to three full-time days of an engineer's time.

42. In addition, implementing the code to search and pull random URLs from Google's index and to search and copy the search queries without identifying information will require extended hours of processing time on Google's computers. Because these are not routine functions for Google, the computers will be processing an additional and burdensome program. Running

such programs above and beyond the normal demand on Google's computers will interfere to some unknown degree with the day-to-day operations of Google. The computers that maintain the index and search queries also process Google's spam filters and process advertising reports. Accordingly, these functions could be slowed down or completely interrupted by processing the Government's request, resulting in lower quality of service to users of Google's search engine and to Google's advertisers.

I DECLARE UNDER PENALTY OF PERJURY that the foregoing is true and correct.

DATED at Mountain View California, this 16th day of February, 2006.

Matt Cutts

MATT CUTTS