PETER D. KEISLER
Assistant Attorney General
THEODORE HIRT
Assistant Branch Director
JOEL McELVAIN, D.C. Bar No. 448431
Trial Attorney
U.S. Department of Justice
Civil Division, Federal Programs Branch
20 Massachusetts Ave., NW
Washington, DC 20001
Telephone: (202) 514-2988
Fax: (202) 616-8202
Email: Joel.L.McElvain@usdoj.gov

Attorneys for Alberto R. Gonzales

IN THE UNITED STATES DISTRICT COURT
FOR THE NORTHERN DISTRICT OF CALIFORNIA
(SAN JOSE DIVISION)

ALBERTO R. GONZALES, in his official capacity as ATTORNEY GENERAL OF THE UNITED STATES,	
Movant,	Case No. 5:06-mc-80006-JW
v.	**Notice of Filing of Supplemental Declaration of Philip B. Stark, Ph.D.**
GOOGLE INC.,	
Respondent.	

NOTICE is hereby given of the filing by Alberto R. Gonzales, in his official capacity as Attorney General of the United States, of the attached Supplemental Declaration of Philip B. Stark, Ph.D. This declaration is filed in support of the Motion to Compel Compliance with Subpoena Duces Tecum, in which the Attorney General seeks to compel Google Inc. to comply with the subpoena issued to it pursuant to Federal Rule of Civil Procedure 45.

Dated: February 24, 2006

Respectfully submitted,

PETER D. KEISLER
Assistant Attorney General

THEODORE HIRT
Assistant Branch Director

/s/
JOEL McELVAIN
Trial Attorney
United States Department of Justice
Civil Division, Federal Programs Branch
20 Massachusetts Ave., NW, Room 7130
Washington, D.C. 20001
Telephone: (202) 514-2988
Fax: (202) 616-8202
Email: Joel.L.McElvain@usdoj.gov

Attorneys for the Movant, Alberto R. Gonzales

SUPPLEMENTAL DECLARATION OF PHILIP B. STARK, PH.D.

1. I submitted a declaration in this matter on 16 January 2006. I am Professor of Statistics at the University of California, Berkeley, where I have been on the faculty since 1988, and where I have held a Miller Research Professorship and have been a Dodson Fellow and a Presidential Chair Fellow. I received a Bachelor's degree from Princeton University in 1980 and a Ph.D. from the University of California, San Diego, in 1986. I was a Presidential Young Investigator and a National Science Foundation Postdoctoral Fellow in Mathematical Sciences. I have been on the editorial board of several journals. I have written over 65 articles and technical reports and an introductory statistics textbook. I have given roughly 130 invited lectures at conferences and universities in 16 countries. I have testified to the U.S. House of Representatives Subcommittee on the Census and the California Senate Natural Resources Committee. I have consulted for the U.S. Department of Justice, the Federal Trade Commission, the U.S. Department of Agriculture, the U.S. Census Bureau, the U.S. Attorney's Office of the Northern District of California, the U.S. Department of Veterans Affairs, the Los Angeles County Superior Court, the National Solar Observatory, public utilities, major corporations, and numerous law firms. Some of my consulting and research relates to the Internet, including characterizing and predicting online consumer behavior and developing search algorithms. I have

been an expert witness or non-testifying expert in cases involving antitrust, employment discrimination, equal protection, fairness in lending, federal legislation, insurance, product liability, intellectual property, trade secrets, truth in advertising, wage and hour disputes, and other matters. A recent version of my CV was attached to the Declaration of Ashok Ramani, dated 16 February 2006, filed by Google in this action.

2. As part of its defense of the Child Online Protection Act (COPA) in ACLU v. Gonzales, Civ. Action No. 98-5591 (E.D. Pa.), the government seeks to determine whether COPA is more effective than content filters at blocking sexually explicit material on the Internet from view by minors. The government has commissioned a scientific study of this question ("the study"). I am involved in the study. The study will compare the effectiveness of filters and COPA by seeing how they perform on a collection of websites that have been categorized by content. To construct a representative set of websites, the U.S. Department of Justice issued subpoenas to several search providers, including Google. The subpoenas asked for a list of URLs from the indexes maintained by the search engines and for a list of queries users had executed.

3. The study includes the following steps. A human being will browse a random sample of 5,000–10,000 URLs from Google's index and categorize those sites by

content. A random sample of approximately 1,000 Google queries from a one-week period will be run through the Google search engine.[1] A human being will browse the top URLs returned by each search and categorize the sites by content. The sites will be used to estimate the fraction of sites in Google's index that contains sexually explicit material and the fraction of queries that returns sexually explicit material. The sites from Google and other sources will also be used to measure how well commercial content filters block sexually explicit material.

4. Google questions the relevance of the data requested in the subpoena. Opposition to the Government's Motion to Compel, 17 February 2006, at 5 (*Opposition*). Google claims to be "the world's most-used search engine."[2] An estimated 25% of Internet searches seek pornography.[3] Over the last few months, about 40% of the 100 most popular search terms appear to seek pornography.[4] Arguably, "Google is one of the largest single gateways on the Internet for pornography."[5]

[1] The government is requesting 50,000 URLs and 5,000 queries even though the study calls for fewer data. Extra data are needed in case of problems with data quality.

[2] *Opposition* at 2.

[3] Vise, D.A., and Malseed, M., 2005. *The Google Story*, Bantam Dell, NY, at 165, citing a 2004 study by Family Safe Media; also see http://www.familysafemedia.com/pornography_statistics.html

[4] *Wordtracker top 1,000 list*, 22 February 2006. Wordtracker is a private company that exists "to answer a fundamental question in the search engine industry: 'What are people searching for on the Web?'" http://www.wordtracker.com/about.html. Wordtracker publishes weekly estimates of the popularity of search terms. Google seems to recommend Wordtracker. Declaration of Matt Cutts, 16 February 2006, at 13.

[5] Vise and Malseed, *loc. cit.*

5. Google notes that it is not the only source of URLs and queries.[6] *Opposition* at 14–15. The study will examine data from a variety of sources. There are sound statistical reasons for using data from a number of sources. For example, the sources might sample different parts of a population[7] or might have different biases. In such cases, increasing the number or variety of sources can improve accuracy more than increasing the number of data from a given set of sources. Google argues that its data cannot be important because the government did not seek data from AskJeeves. *Opposition* at 14. It is particularly helpful to have data from sources that represent a large fraction of the population under study. And according to recent estimates, Google has over 45% of the search market, while AskJeeves has less than 3%.[8] Google also claims that the government already has enough data and that the government has not shown that sample data can prove any fact reliably. *Opposition* at 14. The field of Statistics teaches how to make reliable

[6] For example, Google suggests that the government could draw a sample of URLs from Alexa. *Opposition* at 15. But a sample from the Alexa index need not be representative of Google's index. Alexa is not a search engine: Its searches are "powered by Google." http://alexa.com.

[7] "Population" is a technical term in Statistics. It refers to the collection of things under study—not necessarily to a group of people. The study will sample a number of populations, including the population of sites indexed by search engines and the population of sites search queries return.

[8] Nielsen/NetRatings, 2005. *Volume of search queries jumps 15 percent in past five months, driven by the 'Big Three' search engines, according to Nielsen/NetRatings*, Report pr_051213.pdf. For U.S. searches, comScore estimates Google has 39.8% of the market and AskJeeves has 6.5%. comScore, 2006. *Comscore Releases November Search Engine Rankings,* http://www.comscore.com/press/release.asp?id=694

inferences about a population[9] from sample data. Inferences from a sample are most reliable when the sample draws from the study population as a whole. Google's index and query results reflect a large cross-section of Internet searching that might differ from that of other search engines.

6. Google argues the contrary. The argument contradicts itself. Google claims data from other sources could be substituted for its data. *Opposition* at 14–15. But Google also claims its data are unique and proprietary, constituting trade secrets its competitors could exploit. *Opposition* at 1, 9–11. Both claims cannot be true simultaneously:

Either

i. Data from other search engines make Google's data redundant. Then Google's data are not unique and Google's competitors would not benefit from disclosure.

Or

ii. Google's data are unique. Then data from other search engines are not a good substitute.

There are other contradictions. For example, Google claims that from a sample of queries, one could deduce its users' demographics. *Opposition* at 11. But Google also claims the queries "will not reveal whether the search query was run by a

[9] See footnote 7.

minor or adult, human or non-human, or on behalf of an individual or business. No conclusion can accurately be drawn from this data about individual behavior." *Opposition* at 7. If it is impossible to tell whether a human or non-human generated a query, it is hard to understand how queries could contain reliable demographic information.

7. Google claims it is impossible to use the requested data in a way that reflects realistic search results unless one knows the inner workings of Google's "proprietary and confidential methodologies." *Opposition* at 6. "[U]nless you know *how* Google works, you cannot possibly know *what* Google will return in response to any query. Any assumption to the contrary would be inadmissible speculation." *Opposition* at 8. Speculation is not needed. It is easy to find out what Google will return in response to any query: Just type the query into Google. Millions of people know what Google returns in response to their queries without knowing how Google works. All the study needs from Google is a random sample of URLs from Google's index, a random sample of Google queries, and the results of running about a thousand of those queries through Google's search engine.

8. Google notes that the presence of a URL in its index does not indicate how often the URL is returned by searches. *Opposition* at 8. That is why the government requested both a random sample of URLs from Google's index and a

sample of Google queries. The sample of URLs will be used to estimate aggregate properties of the sites Google has indexed. The sample of queries will be used to estimate aggregate properties of sites returned by Google queries.

9. Google notes that a site's URL does not necessarily show whether the site contains sexually explicit material. *Opposition* at 8. In the study, a human being will view each site in the sample—not merely its URL—to categorize the site's content.

10. Google claims a sample of queries would reveal "the number of queries that Google can or does process, its capabilities of processing certain lengths and types of queries, its market share in the United States and other countries, and even the demographics of its users." *Opposition* at 10. That Google receives a query does not mean Google processes the query in any particular way.[10] Nor can I think of a reliable way to deduce demographic information from a sample of queries.[11] True, the number of queries Google receives in a given week—together with the number of queries all other search engines received that week—allow one to calculate Google's market share that week. But estimates of Google's market share are

[10] If you send me a message in Finnish, that does not mean that I can read Finnish.

[11] To illustrate, I might search for "BMW" because I am a 45-year-old female attorney who owns a BMW automobile, because I am a 16-year-old male student who dreams of owning a BMW motorcycle but cannot afford one, because I am a 30-year-old male bicyclist who just collided with a parked BMW, or because I am a 65-year-old female investor curious about Google's recent dispute with BMW. See also paragraph 6.

readily available, for example, in reports by Nielsen/NetRatings and comScore.[12]

11. Google claims that from a random sample of its index "one could estimate, among other things, the size of Google's index; how deeply Google crawls in different countries or languages; and the ability of Google's crawl metrics to measure the reputation of pages or domains." *Opposition* at 10–11. A random sample of 1 million URLs from Google's index reveals that the index contains at least 1 million URLs, but nothing else about the size of the index. Until late September, 2005, Google gave the size of its index on its homepage: The last count was 8.17 billion pages.[13] It is not clear whether "how deeply Google crawls" means "folder depth" or "click depth"[14] or something else. I cannot think of a way to estimate the click depth of Google's crawling from a sample of URLs without knowing which other pages in Google's index have links to those URLs. From a sample it might be possible to estimate the folder depth to which Google crawls some parts of the web. But both the click depth and folder depth of Google's

[12] See footnote 8, paragraph 12, and footnote 17. The study will not analyze the fraction of queries in different languages. To infer a query's country of origin from its language seems tenuous. Estimates of Google's international market share are available from other sources, for example, from comScore. ComScore, 2004. *More Than 40 Million Consumers in the U.K., France and Germany Used Search Engines in April, According to comScore Networks.* http://www.comscore.com/press/release.asp?id=464

[13] http://googleblog.blogspot.com/2005/09/we-wanted-something-special-for-our.html Liedtke, M., 2005. Google to take down front-page boast about index size, *USA Today*, 27 September 2005. http://www.usatoday.com/tech/news/2005-09-27-google-count_x.htm

[14] Folder depth is essentially the number of slashes in the URL after the domain name. Click depth is essentially the number of clicks it takes to get to the URL.

crawling can be estimated by running Google searches restricted to the domain in question.[15] A random sample of 50,000 URLs from Google's index would be adequate for the study. Any estimates of folder depth in different domains from a sample of that size would have large uncertainties. Most domains would not appear in the sample at all. Few domains, if any, would appear more than once. I do not believe it is possible to estimate "the ability of Google's crawl metrics to measure the reputation of pages" from a sample of URLs alone.

12. Google claims search queries can contain personally identifying information. The government requested that Google remove any identifying information. The text of a query does not identify the user: There is no reliable way to tell who did the typing from what was typed.[16] A search for my name, address, and social security number might have been run by me or by someone trying to learn about me. A snippet of a document could have been pasted into a search box accidentally by the document's author or by someone reading or editing or even plagiarizing the

[15] Google searches can be restricted to specific domains, and can be restricted to look for text only in the URL of the page. http://www.google.com/help/refinesearch.html One can find many or all of the pages Google has indexed in a given domain by searching for the domain name in the URL, with the domain restricted to the domain. For example, the query "allinurl:bmw.com site:bmw.com" shows that Google has crawled about 600 pages in the bmw.com domain. The query "engine site:bmw.com" shows that Google has crawled the bmw.com site to a folder depth of at least 13, because the search results include https://b2bpapp6.bmw.com/public/en/b2b/entwicklung/pap/apps/online/prisma/doc/fm/dokument/dokument_auswahl/intro.html.

[16] See, for example, Liptak, A., 2006. In Case About Google's Secrets, Yours Are Safe. *The New York Times*, 26 January 2006.

document. The study does not involve examining the queries in more than a cursory way.[17] It involves running a random sample of the queries through the Google search engine and categorizing the results.

13. Google queries are disclosed routinely to third parties when a user clicks any link in Google search results.[18] The government seeks less information about queries than Google publishes in Google Zeitgeist.[19]

14. Google worries that I might reveal its confidential information. *Opposition* at 13. I am often privy to confidential information and data as a consultant and expert witness. I have signed numerous non-disclosure agreements and declarations attesting that I will obey protective orders. And I have abided by them. I signed a

[17] Some quality control checks will be performed.

[18] Unless the user has disabled "referrer logging," the browser reports the URL it came from when a link is clicked. The URL for Google search results contains the query string. For example, the URL of the page of results that the Google search engine returns when I search for "google subpoena" is http://www.google.com/search?client=opera&rls=en&q=google+subpoena&sourceid=opera&ie=utf-8&oe=utf-8. This URL shows that I use the Opera browser, that I searched for "google subpoena," and other parameters of my search. If I clicked on any link in the search results, all that information would be transmitted to the website that is the target of the link. Google encourages advertisers to use this information to track traffic: "Web server logs provide electronic recordings of where your website traffic originates. To open the log file in your web server software, use a text editor such as Notepad. This log file has an entry for each click to your site. To see how many clicks come from a particular source, just count the entries where the source (such as 'Google' or Google Network sites and products) appears in the referring URL." https://adwords.google.com/support/bin/answer.py?answer=6429&topic=35

[19] Google Zeitgeist (http://www.google.com/zeitgeist) lists the most popular Google queries overall, and in various categories, for various time periods.

declaration that I will obey the protective order issued in this case by the Eastern District of Pennsylvania. A copy of that declaration is attached as Exhibit A. Google's "deep[] concern[]"[20] about my involvement with Cogit.com is especially misplaced: Cogit.com went out of business several years ago.[21]

15. Google says it would take three to eight days of engineering time to sample their URLs and queries. *Opposition* at 16. That estimate seems high because:

i. Google acknowledges it has a query log and an index. *Opposition* at 11. Declaration of Matt Cutts, 16 February 2006, at 11–12.

ii. Google has produced samples from its index before. *Opposition* at 11.

iii. Google has tools to generate reports from its query logs, for example, Google Zeitgeist.[22] Generating a random sample would likely involve only a minor modification of such reporting tools.

iv. In response to the government's subpoena, other search providers produced electronic files containing random samples of 1 million URLs and a full week of user queries, without complaint.

16. Google claims it would take months to negotiate an agreement with the

[20] *Opposition* at 13.

[21] The website http://www.cogit.com is registered to Collins Technologies, Inc., according to a "whois" lookup on 24 February 2006. The site contains a definition of "web analytics" and advertisements by Google.

[22] http://www.google.com/zeitgeist

government on how to define a "random" sample. *Opposition* at 17. In Statistics, as in many fields, it is possible to find experts who disagree on fine philosophical or technical points, but I cannot think of a statistician who would dispute how to draw a simple random sample. Google can draw the samples using standard methods the government considers adequate.[23]

17. Google claims that a particular script written for the Firefox browser will skew Google's query logs by adding random searches for pornography. *Opposition* at 7. But the number of people who use the Firefox browser and installed the script is certainly tiny compared to the number of Google users. The script just adds an extra query each time one of those people queries Google. Google processes hundreds of millions of queries per day.[24] It is extremely implausible that the script adds an appreciable number of queries. Moreover, the three queries in the script might be its author's idea of political humor, but they are not searches for pornography.[25] I ran the three queries on Google on 22 February 2006. None returned pornography in the first page of results.

[23] A suitable method from a classic text on sampling, such as Cochran, W.G., 2002. *Sampling Techniques*, *3rd edition*, John Wiley & Sons, NY, or Kish, L., 1965. *Survey Sampling*, John Wiley & Sons, NY, would be adequate for the study.

[24] Hursh, P., 2006. Marketing in the Search Tail: Is the Pain Worth the Gain? *Search Engine Watch.* http://searchenginewatch.com/searchday/article.php/3579396
Vise and Malseed, *op. cit.*, at 5.
Malone, M.S., 2003. Inside the Soul of the Web, *Wired*, 11.05.
http://www.wired.com/wired/archive/11.05/google.html

[25] http://www.rungie.com/~craig/DoJ.user.js (accessed 23 February 2006)

18. Google points out a number of issues with the data the government has requested. I have considered them all;[26] some are discussed above. The goal of the study is to estimate aggregate characteristics of websites indexed by search engines and websites returned by searches and to compare how well content filters and COPA block sexually explicit material. For those ends, I believe the study is sound and that the data requested from Google—in conjunction with data obtained elsewhere—will lead to reliable conclusions.

I declare under penalty of perjury that the foregoing is true and correct.

______________________________ Dated 24 February 2006.

Philip B. Stark

[26] For example, for the study, it is not essential to know whether a particular query was run by a child or an adult; by an individual acting in a personal or business capacity; or by a person or by a software "bot" on a person's behalf. Some Google users customize some parameters of their searches (*Opposition* at 6), but running a random sample of queries through Google's search engine with default parameter settings is still informative. Google's algorithms change (*Opposition* at 7), but running a random sample of queries still gives an estimate of the fraction of queries that returns sexually explicit material when the queries are executed, and still provides a set of sites that can be used to test content filters. A site's content might have changed since Google indexed it (*Opposition* at 8), but its content when viewed is what matters for the study.

EXHIBIT A

IN THE UNITED STATES DISTRICT COURT
FOR THE EASTERN DISTRICT OF PENNSYLVANIA

AMERICAN CIVIL LIBERTIES UNION, et al.,	)	
Plaintiffs.	)	
v.	)	Civil Action No. 98-CV-5591
ALBERTO R. GONZALES, in his official capacity as Attorney General of the United States,	)	
Defendant.	)	

DECLARATION

The undersigned hereby declares under penalty of perjury that he (she) has read the Agreed Protective Order (the "Order") entered in the United States District Court for the Eastern District of Pennsylvania in the above-captioned action, understands its terms and agrees to be bound by each of those terms. Specifically, and without limitation, the undersigned agrees not to use or disclose any confidential information made available to him (her) other than in strict compliance with the Order.

DATED: 9/14/2005

BY: Philip B. Stark
(type or print name)

CERTIFICATE OF SERVICE

I hereby certify that I have made service of the foregoing Notice of the Filing of the Supplemental Declaration of Philip B. Stark, Ph. D., by depositing in Federal Express at Washington, D.C., on February 24, 2006, true, exact copies thereof, enclosed in an envelope with postage thereon prepaid, addressed to:

Albert Gidari, Jr., Esquire
Perkins Coie, LLP
1201 Third Avenue
Seattle, Washington 98101-3099
(Counsel for Respondent Google Inc.)

Lisa Delehunt, Esquire
Perkins Coie, LLP
180 Townsend Street, Third Floor
San Francisco, CA 94107
(Counsel for Respondent Google Inc.)

Aden J. Fine, Esquire
American Civil Liberties Union Foundation
125 Broad Street
New York, New York 10004
(Counsel for Plaintiffs, *ACLU v. Gonzalez*, E.D. Pa. No. 98-cv-5591)

/s/
JOEL McELVAIN
Attorney