Hastings Science & Technology Law Journal

Search Query Privacy: The Problem of Anonymization

by RON A. DOLIN, J.D., PH.D.

CITE AS: 2 HASTINGS SCI. AND TECH. L.J. 137

Abstract

Search queries may reveal quite sensitive information about the querier. Even though many queries are not directly associated with a particular person, it has been argued that the IP addresses and cookies of the users can often be sufficient to figure out who the querier is, especially if tied to information from ISPs regarding IP address assignments at the time of the relevant query. Given that the queries have been subject to discovery both by various governments and third parties, there has been great concern for how to keep such queries private. A typical approach to such privacy legislation, especially in Europe, has been to require either destruction of the data so that it is no longer available for discovery, or anonymization so that it cannot be associated with a particular person. This solution has never been proposed for personal data such as medical information used by doctors or financial information used by credit agencies. Instead, there seems to be an assumption about these types of data that their long-term storage is necessary and/or beneficial to the individual associated with them, or at least to society at large. The framework for maintaining the privacy of these data turns on safeguards where it is being held, user control of its retention and accuracy, and strict legal limitations regarding its discovery. This article briefly reviews a few legal frameworks for data protection both in the U.S. and in Europe. It presents several arguments that the deletion or anonymization of search query data is problematic, and describes a framework similar to the way we handle health data that is more beneficial to all stakeholders. Such an approach would lead to a more uniform solution to data protection in which maintaining search query privacy would not sacrifice the benefits of long term, confidential storage of the data.

Search Query Privacy: The Problem of Anonymization

by RON A. DOLIN, J.D., PH.D.[*]

I. Introduction

A. The Nature of the Problem

Search queries may reveal quite sensitive information about the querier. One can imagine the potentially compromising nature of queries and result clicks: a spouse looking up STD's; a student seeking free copyrighted music or video downloads; someone inquiring about nuclear bomb or other WMD technology; a citizen posing questions about a political group within a country that disfavors or forbids it. Even though most queries are not directly associated with a particular person, corresponding identifying information can often be sufficient to figure out who the querier is, which can create a trail of sensitive information.

While most search engines have policies that protect users' privacy to some degree,[1] search queries and other user-generated content have been the subject of governmental, private, and international discovery.[2] As a result, many countries have initiated

[*] Ph.D. in Computer Science. Disclosure: the author worked several years at Google as an engineer before leaving for law school, and maintains an interest in the company; the viewpoints here are the author's, not necessarily shared by Google. The author would like to thank Prof. Robin Feldman, Prof. Mark Lemley, and Frank Cusack for their valuable feedback on this work. An extended abstract of this article appeared in the *Proceedings of the Intelligent Privacy Management Symposium*, TECHNICAL REPORT SS-10-05 1, 59–65 (*Privacy 2010 Symposium*, March 22–24, 2010, Stanford University, California).

[1] *See*, *e.g.*, Google.com, Privacy Center, http://www.google.com/intl/en/privacy.html (last visited Apr. 23, 2010).

[2] For example, after the Federal Courts struck down the first version of the Child Online Protection Act in 2006, the U.S. Department of Justice sought search queries from the four leading search engines in order to establish the percentage of searches related to pornography – only Google fought the subpoena. *See*, *e.g.*, American Civil Liberties Union v. Gonzales, 478 F.Supp.2d 775 (E.D. Penn. 2007); Gonzales v. Google, Inc., 234

policies that seek to protect these queries. In particular, there has been a strong push to force search engines to delete and/or anonymize these data after a few months so that they are not available for later discovery or other uses. However, these data can be quite useful for many reasons, and some search engines have been reluctant to follow this method of privacy protection.[3]

When a user enters search terms in an internet search engine, the query terms are logged by the search engine. In addition to the query terms, the log data includes items such as the type and version of the user's web browser, IP address, and various "cookie" information. Cookies, viewable in most web browsers within "options" or "preferences", are used to allow the search engine to keep track of some information associated with a user that is not sent by the web browser with the query, such as prior queries issued from the same browser, the user's email account if signed in, etc. Anonymization generally consists of deleting cookie information and either completely or partially removing the IP address such that it can not be traced back to an individual machine.[4]

There is a growing demand in the U.S. for search query anonymization, as discussed in a New York Times article from 2008:

F.R.D. 674 (N.D. Cal. 2006); Wikipedia.com, Child Online Protection Act Overview, http://en.wikipedia.org/wiki/Child_Online_Protection_Act (last visited Apr. 23, 2010); GoogleBlog, *Judge tells DoJ "No" on search queries*, http://googleblog.blogspot.com/2006/03/judge-tells-doj-no-on-search-queries.html (last visited Apr. 23, 2010). In another well-known case, Viacom was able to obtain YouTube queries in an attempt to argue that the YouTube website is used pervasively for illegal music and video downloads, although both sides agreed to anonymize the queries prior to the data being handed over due to the ensuing public outcry. *See, e.g.,* Miguel Helft, *Google Told to Turn Over User Data of YouTube*, N.Y. TIMES, July 4, 2008, at C4, *available at* http://www.nytimes.com/2008/07/04/technology/04youtube.html (last visited Apr. 23, 2010). Finally, in another (in)famous case, Yahoo data was used by the Chinese government to identify and convict a journalist in 2004. (Yahoo released the identity of the holder of an email account – not search related). *See*, *e.g.*, Zachary Coile, *Lawmakers Blast Yahoo Executives for Helping China Jail Dissident*, S.F. CHRON., Nov. 7, 2007, at A-1, *available at* http://www.sfgate.com/cgi-bin/article.cgi?f=/c/a/2007/11/07/MN2NT7C99. DTL (last visited April 23, 2010).

[3] *See*, *e.g.*, JBE, *Google in dispute with EU data protectionists about retention of IP addresses*, THE H, Feb. 26, 2008, http://www.h-online.com/news/Google-in-dispute-with-EU-data-protectionists-about-retention-of-IP-addresses--/110192 (last visited Apr. 23, 2010); TRK, *Google resists further regulation on retention of search query data*, THE H, Apr. 9, 2008, http://www.h-online.com/news/Google-resists-further-regulation-on-retention-of-search-query-data--/110501 (articles from 2008 discussing the differences of opinion between Google and the EU regarding data retention, anonymization, and the privacy implications of IP addresses).

[4] *See* discussion in Section VI, *infra* (The anonymization of zip codes by removing the last 2 or 3 digits is similar to the anonymization discussed here).

> Yahoo's new data retention policy is the most restrictive among major search engines in the United States and will most likely put pressure on rivals like Google and Microsoft to shorten the time they keep information about their users.
>
> It comes at a time when some privacy advocates are planning a renewed push for legislation that would regulate the data retention and online advertising practices of Internet companies, which they say has a stronger chance of passing with a new Congress and president in Washington.
>
> Already Representative Edward J. Markey, a Massachusetts Democrat who is chairman of the House Subcommittee on Telecommunications and the Internet, praised Yahoo for setting a new privacy standard.[5]
>
> "I urge other leading online companies to match or beat the commitments announced by Yahoo," Mr. Markey said in a press release.[6]

Privacy, however, is not the only issue one should consider here, nor is anonymization necessarily the right solution to the set of problems faced. This article sets out several issues suggesting that the push toward anonymization and deletion is ill-conceived. This article also suggests borrowing from, and adding to, other data protection schemes, such as that used for health records, that serve to maintain long-term information. Such an approach seeks to enforce privacy through data protection, balancing privacy concerns with other factors such as the long-term usefulness of the data and problems associated with data destruction.

B. Overview

As background, Section VI discusses several types of existing data protection referred to in the main sections. First, it introduces several schemes used in the U.S. The U.S. approach has been somewhat piecemeal, as evidenced by the different acts and requirements for different types of data (e.g., census, health, financial,

[5] Representative Markey is no longer the chairman of the committee at the time of this writing.

[6] Miguel Helft, *Yahoo Limits Retention of Search Data*, N.Y. TIMES, Dec. 17, 2008, at B3, *available at* http://www.nytimes.com/2008/12/18/technology/internet/18yahoo.html (last visited Apr. 23, 2010).

and communications data). Each data type involves different protection schemes, with some kept completely confidential (e.g., census data), some allowing for criminal prosecution if leaked (e.g., census and health data), some requiring probable cause and a search warrant (e.g., electronic communications in transit), and some being permitted to be sold to third parties unless the individual "opts out" (e.g., financial data).

Within the E.U., there is a more comprehensive approach established under Directive 95/46/EC, on "the protection of individuals with regard to the processing of personal data and on the free movement of such data."[7] It is often referred to as the "data protection" directive. However, since directives are not self-executing, each E.U. member country has implemented the Directive differently.[8] While the details of the various implementations are beyond the scope of this article, the Directive itself establishes an important framework that is applicable to all forms of data.[9] In particular, the Directive established an ongoing advisory committee, the so-called Article 29 Working Party (established by the Directive's Article 29), which is tasked with dealing with ongoing issues of specific types of data, resolving ambiguities, and so forth.[10] In 2008, the Working Party released WP148, its "Opinion 1/2008 on data protection issues related to search engines."[11] Given the comprehensive nature of that document, it provides a convenient and comprehensive perspective arguing for search query data anonymization.

Using WP148 as a backdrop, Section II presents several arguments for keeping the search query data intact instead of

[7] Council Directive 95/46, 1995 O.J. (L 281) (EC) 1, 1.

[8] *See*, *e.g.*, Peter Fleischer, *Lead Data Protection Authority*, http://peterfleischer.blogspot.com/2009/02/lead-data-protection-authority.html (last visited Apr. 23, 2010).

[9] *See*, *e.g.*, Peter Fleischer, *Launching another "global" forum to talk about privacy*, http://peterfleischer.blogspot.com/2009/01/launching-another-global-forum-to-talk.html (last visited Apr. 23, 2010) and 30th Int'l Conference of Data Protection & Privacy Comm'rs, *Resolution on the urgent need for protecting privacy in a borderless world, and for reaching a Joint Proposal for setting International Standards on Privacy and Personal Data Protection*, (Oct. 17, 2008), *available at* http://www.lda.brandenburg.de/sixcms/media.php/3509/resolution_international_standards_en.pdf (last visited Apr. 23, 2010 (discussing other international efforts to harmonize data protection regulations).

[10] Council Directive 95/46, art. 29, 1995 O.J. (L 281) (EC) ("Working Party on the Protection of Individuals with regard to the Processing of Personal Data" establishes the independent committee with advisory status).

[11] The Data Protection Working Party, *Opinion 1/2008 on Data Protection Issues Related to Search Engines*, WP148, 00737/EN (Apr. 4, 2008).

anonymizing them. The list includes factors such as the potential usefulness of the data; exceptions such as criminal investigations; registered vs. non-registered use; the purpose for the data collection; and data ownership. These arguments serve as counterweight to the orientation of WP148, which focuses almost exclusively on privacy. This focus may lead one to conclude that deletion is the best solution, and the point of Section II is to highlight that such a solution comes at a cost. By viewing anonymization as half of a trade-off, we can step back and see if the protection approaches we take for other forms of data might be adequate to protect search query data, and, if so, allow us to simultaneously protect privacy while benefiting from the continued use of the data. Section III briefly outlines one such approach, borrowing from other data protection schemes in an attempt to delineate boundaries such as which stakeholders might be responsible for which roles. That is, what might be considered the proper role for governments, business development, engineering, and users, in order to maximize the protection/use trade-off. Section IV argues that in one way or another, each stakeholder has contributed to the call for anonymization, potentially against their own interests, and what they might do differently to safely allow for the realization of the vast hidden potential of the data's long term storage.

II. Anonymization Problems

The universal reason to anonymize search engine query logs, given both domestically and internationally, is to protect the privacy of the users. Perhaps the most detailed argument given from that vantage point is the Article 29 Working Party report WP148, discussed *infra*. However, there are many arguments that run counter to that perspective. In presenting some of them, I do not attempt to stay within a framework of the law of the United States. Although in some cases American case law is on point and helps to delineate the issues, other cases are more aligned with a European perspective. I use WP148 as a framework to present anonymization arguments due to its comprehensive nature, not because it is a European approach.[12] Taken as a whole, the arguments reinforce each other in such a way that they highlight the problems with anonymization and bring to light the trade-offs that are being made in the name of privacy.

[12] In addition, the Working Party is derived from Council Directive 95/46, 1995 O.J. (L 281) (EC), which is not self-executing. The details of the implementation of that Directive, quite varying from country to country, is beyond the scope of this article.

The following arguments are an attempt to present one side – that of retaining the information. The text often uses interchangeably the terms anonymization and deletion. The reason is that anonymization of search query data entails deletion of certain pieces of information associated with the actual search query terms, such as IP addresses.

A. Data Usefulness

Enforcing privacy through anonymization comes at the cost of losing any benefit derived from the use of non-anonymized data. In a nutshell, it is difficult to overstate the vast number of potential uses for search query information, which are limited only by one's imagination. The data are not only valuable to the user individually, but also to the search engine, the government, and society at large. Uses include the improvement of current tools, the development of new ones, the predictive power through data analysis, and the compelling historical, statistical, and scientific potential down the road.

For an individual, a registered search history can be an aid to search results and help reduce irrelevant advertising. It can help differentiate ambiguous terms on an individual basis (e.g., jaguar – car vs. cat); help with personalized spell corrections and term substitution; indicate which languages someone has used; and aid in determining appropriate levels of filtering for profanity, sexual content, etc.

Search query data are used to improve the search algorithm, as a defense against "malicious access and exploitation attempts," to ensure system integrity (e.g., preventing click fraud), and to protect users (e.g., preventing spam, phishing, etc.).[13] IP addresses and cookies are important in all of these. Here is how Google has described some of the benefits of using this data:

> 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

[13] Youtube, Peter Fleischer on Privacy, http://www.youtube.com/watch?v=JNu1OtkWrOY (last visited Apr. 23, 2010). For more details *see* Google, *Using data to help prevent fraud*, http://googleblog.blogspot.com/2008/03/using-data-to-help-prevent-fraud.html (last visited Apr. 23, 2010); Google, *Using log data to help keep you safe*, http://googleblog.blogspot.com/2008/03/using-log-data-to-help-keep-you-safe.html (last visited Apr. 23, 2010).

> 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.
>
> . . .
>
> 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 . . .).[14]

For society at large, consider the following example from Google. Flutrends detects regional flu outbreaks two weeks before similar detection by the CDC by analyzing flu-related search terms such as "flu" or "influenza" in coordination with the geographical region as determined by IP addresses (though the data are anonymized prior to release in a similar way to census data).[15] The work is accurate enough to have warranted publication in the well-known science journal Nature.[16] As discussed in Section II(D), *infra*, IP addresses can frequently yield geographical information down to the city or zip code level without identifying a user's identity. In principle, knowledge of a coming local flu outbreak can give people advance notice to get flu shots, wash hands more, etc., which can save lives. This tool was developed by looking over 5 years worth of non-anonymized data.

Another example, perhaps just for curiosity, is what people are searching for in your local area.[17] However, imagine looking for a

[14] Paul Haahr & Steve Baker, Official Google Blog, *Making search better in Catalonia, Estonia, and everywhere else*, Mar. 25, 2008, http://googlepublicpolicy.blogspot.com/2008/03/making-search-better-in-catalonia.html (last visited April 23, 2010).

[15] Google.org, *Explore flu trends around the world*, http://www.google.org/flutrends/; *see* Section VI(A)(2), *infra.*

[16] Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski & Larry Brilliant, *Detecting Influenza Epidemics Using Search Engine Query Data*, 457 NATURE, 1012 (2009).

[17] *See*, *e.g.*, San Francisco within last 7 days: http://www.google.com/insights/search/# (select "Locations" under "Compare by:"; then select "United States"

signal that would indicate pending economic problems, such as a rise in several regions' queries about foreclosures or bankruptcies, and imagine detecting that signal early enough to prevent a national or international economic crisis. It would require several years worth of data to be able to detect such a signal with sufficient reliability to be able to act on it. How many jobs or retirement funds could potentially be saved?

Imagine comparing the spread of early domesticated plants and animals with the spread of ideas today. In *Guns, Germs, and Steel*, for example, Pulitzer Prize-winner Prof. Jared Diamond concluded that the east/west orientation of Eurasia facilitated a much faster spread of early agriculture than the north/south orientation of the Americas due to the corresponding similarity of climate in the former case as opposed to the latter.[18] What is the online equivalent for the spread of ideas—urban/rural? Do ideas show up in web pages before or after they show up in queries, and how does this map region by region? How do virtual communities map to geographical ones?

If census data tells us who and where we are, then search queries tell us what we're thinking. Imagine what one could study with 100 years of search query data – non-anonymized. The assumption that such data are expendable is questionable at best, and certainly an odd determination to leave to the government; we give up a lot of value by deleting them rather than securely keeping them around.

As discussed in Section VI(A)(3), *infra*, credit information is stored in order to protect lenders and the financial system – that is, the collective. We keep data about an individual with bad credit, to his detriment, in order to facilitate a functional financial system, and to improve its efficiency. The individual who cannot get a loan due to his (accurate) bad credit has no control over the data's storage. In the case of search query data, when used as intended, there is no detriment to anyone in order to benefit the collective. This is similar to our treatment of census data, even though that data collection is compulsory, while for search queries it is voluntary. The trade-off we make here in the name of privacy is the loss of the vast potential usefulness of the data. If they can be kept intact safely, however, the apparent dichotomy goes away.

"California" "San Francisco" under "Locations:"; then select "Last 7 days" under "Filter:").

[18] JARED DIAMOND, GUNS, GERMS, AND STEEL (W.W. Norton 1997).

B. Exceptions to Deletion

Not surprisingly, Directive 95/46/EC, which authorized the creation of the Article 29 Working Party that authored WP148, lists several exceptions to the policy of data deletion. In general, the directive requires that data be collected, used, and/or retained only for "specified, explicit and legitimate purposes."[19] However, exceptions are provided for "processing of data for historical, statistical or scientific purposes . . . provided that Member States provide appropriate safeguards."[20] Other exceptions are found under the Exemptions and Restrictions section, including national security, defense, public security, criminal prosecutions, and even "breaches of ethics for regulated professions."[21] The question here is whether search query data might qualify for any of these exceptions, and, if so, what to do about it.

Arguably, national census data would fall under the exception category given their strong historical, statistical, and scientific significance. As discussed in Section VI(A)(1), U.S. census data date back to colonial times and seem to retain detailed identifying information, consistent with the European framework of allowing such data to be retained given their nature, even though their collection is compulsory.[22] It could be argued that there is important historical, statistical, and scientific significance to search query data, especially if they are associated with geographical location. This would be true even if IP addresses are not reliably identifying – they are still often sufficient to determine a city.[23] The potential usefulness of these data for these purposes is discussed Section II(A), *supra*. The point here, though, is that a reasonable argument could be made that these data qualify as having strong historical, statistical, and scientific significance. To the degree that they do, this would argue against anonymization of the data *in storage*, rather than anonymizing only the *access*.

[19] Council Directive 95/46, art. 6(1)(b), 1995 O.J. (L 281) 31, 48.

[20] *Id.*

[21] Council Directive 95/46, art. 13(1), 1995 O.J. (L 281) 31, 42.

[22] This highlights the distinction between anonymized access and anonymized storage, further discussed *infra* Section III, since, although the identifying information seems to be retained, census data is publicly accessible only in anonymized form (through the Bureau's API, published reports, etc.).

[23] The issue of whether or not IP addresses and cookies are reliably personally identifying is discussed *infra*, dealing with registered vs. non-registered users.

Other potential exceptions for which search query data might qualify are national security and criminal investigations. As mentioned in Section II(D), *infra*, European telecoms and ISP's are mandated by Directive 2006/24/EC to store call and connection data for up to two years under the national security and criminal investigation exceptions of Directive 95/24/EC.[24] The issue for this section is simply whether search query data are likely to have value from such a perspective.[25] I distinguish national security from criminal prosecutions under the assumption that the former is more motivated by crime *prevention* while the latter is more concerned with *procedural safeguards*. Thus, if fishing for information is permitted for national security, I assume that that would not necessarily require the production of admissible evidence for the purposes of criminal prosecution. Under this model, I separate whether search query data might be useful in aiding the detection and prevention of, say, terrorist activities where probable cause may not be available, from the more controlled search and seizure activities associated with criminal investigation and prosecution.[26]

As an example, consider the case of the threatened attack of the U.C. Hastings campus in 2007:

[24] Council Directive 2006/24, 2006 O.J. (L 105) 54 (EC) makes reference in its preamble (4) to "safeguard[ing] national security (i.e. State security), defence, public security or the prevention, investigation, detection and prosecution of criminal offences or of unauthorised use of the electronic communications systems" from Council Directive 2002/58, art. 15(1), 2002 O.J. (L 201) 37, 46 (EC), which in turn cites the exceptions in Council Directive 95/46, art. 13(1), 1995 O.J. (L 281) 31, 42 (EC). for the same reasons.

[25] In particular, a major concern in this regard seems to be the potential for governmental abuse of the data. If we cannot trust the government not to abuse this data, however, what is to stop them from capturing the data in transit when the queries are originally issued, or secretly mandating that the data be forwarded to them for storage elsewhere prior to anonymization? Safeguarding the information against abuse for criminal prosecutions could be facilitated by a legal framework similar to the exclusionary rule for any illegal search or seizure. As will be discussed in Section III, the legal framework could work against using the data for fishing for typical criminal activity, while still making it available for the prevention of terrorist attacks and/or adequate criminal subpoenas under a probable cause model. In any case, it is a somewhat odd argument that we should force the destruction of property not because we are concerned with abuse by the property owner, but with abuse by the government. That is, the government does not trust itself with this data, but it does with census, tax, and health information. As discussed *infra*, companies have an incentive, and perhaps an obligation, to fight spurious subpoenas for search query data, though perhaps they have not come to realize that yet.

[26] I am not trying to argue for or against particular activities that might be employed for national security; rather, I am only discussing whether search query data might be sufficiently useful to such activities to qualify for exception purposes.

> On April 18, 2007, two days after the Virginia Tech massacre, an individual under the moniker "Trustafarian," a first-year law student at UC Berkeley's Boalt Hall School of Law, [anonymously] posted the following message in an AutoAdmit thread titled "Just decided not to do a murder-suicide copycat at Hastings Law":
>
>> Date: April 18th, 2007 1:35 PM
>>
>> Author: Trustafarian
>>
>> I went to bed all set for "Bloody Wednesday," but when I woke—to sun, to flowers in bloom—I just couldn't bring myself to suit up.
>>
>> Maybe tomorrow; I hear rain's in the forecast.
>>
>> (http://www.autoadmit.com/thread.php?thread_id=616215&forum_id=2#7956138)
>
> The posting was later edited by the poster to read "wgwag," (as it currently reads now) which stands for "White Girls With Asian Guys," in an attempt to make it appear as if the original posting were intended as a joke.
>
> Hastings College of Law, acting on the advice of the Federal Bureau of Investigation, cancelled [sic] classes and evacuated the building at 3:22PM.[27]

The website used for this posting, AudoAdmit, does not log IP addresses in order to maintain the anonymity of its users.[28] One could argue that in the case of such postings, any information that might lead to the identity of the poster could help either prevent a potentially tragic act or determine if the posting was a hoax in order to prevent needless disruptions at schools, airports, etc. If search history is relevant in these cases, as one imagines it could be (e.g. searches for weapons, WMDs, airline schedules, etc.), IP addresses and search cookies could help connect the dots.[29] In the case of

[27] Wikipedia, AutoAdmit: U.C. Hastings Evacuation, http://en.wikipedia.org/wiki/AutoAdmit#U.C._Hastings_evacuation (last visited on April 23, 2010) (citations removed).

[28] David Margolick, Portfolio.com: Two Lawyers Fight Cyber Bullying, http://www.portfolio.com/news-markets/national-news/portfolio/2009/02/11/Two-Lawyers-Fight-Cyber-Bullying (last visited Apr. 23, 2010).

[29] AutoAdmit is also involved in a well-known defamation case, Margolick, *supra* note 28. However, since searches by their nature are not posted, the relationship between defamation and anonymized logging is beyond the scope of this article. *See* Doe v. Cioli, 611 F.Supp.2d 216 (D. Conn. 2009); *see also* Wall Street Journal Online, http://online.wsj.com/public/resources/documents/aaComplaint.pdf.

terrorism, the value of this information might not be realized for months or even years. As with the historic, statistical, and scientific value, IP addresses and search cookies are useful even though they are not, in and of themselves, reliably identifying.

Whether for the historic, statistical, or scientific value, or to aid in national security or criminal investigations, access to search query data arguably could be helpful and may well qualify as a valid exception. If so, then these standard exceptions would be consistent with the other arguments that the data should be retained intact.

C. Anonymization is Unnecessary

Assuming that search query data are useful in some way, are there frameworks for protecting intact data so that anonymization is unnecessary? Section III, *infra*, provides one possible framework for protecting intact search query data. It is not by any means the only possibility, as evidenced by the various approaches discussed in Section VI, *infra*. Like health data, it assumes that search query data are sensitive but less desirable to criminals than financial data, and that they will remain in long-term storage. Like census data, it differentiates between anonymous storage and anonymous access. Like electronic communications, it assumes that a warrant should be required for unauthorized access. It proposes criminal liability for knowingly giving data to unauthorized persons. It even suggests that the "identifying" information be physically and logically segregated from the rest of the data. The point is, though, that many forms of data protection are available that would serve to minimize the possibility of unauthorized disclosure while allowing for the benefits of improved service, innovation, and an invaluable historical record. By our defining a reasonable data protection standard on search query data, search engines could keep the data if they view them as sufficiently valuable to warrant the protection costs. That is likely a proper boundary between governmental regulation and market forces.

D. The Many Faces of IP Addresses – Registered vs. Non-registered Users

Another reason that anonymization is unnecessary is that *actual* identification is left as an option to the user. IP addresses and cookies are less revealing than many privacy advocates acknowledge. They become reliably identifying only in cases of a registration, something that users can choose to avoid by performing searches without first logging in.

To clear out or correct search histories – for non-registered users – WP148 states that users can validate that they created the searches via ownership of the relevant IP address, by showing proof from their access provider that they owned the relevant IP address at the time of the search:

> Search engines should respect the rights of data subjects to access, and where appropriate to correct or delete information held about them. These rights apply foremost to the data from authenticated users stored by search engines, including personal profiles. However, *these rights also apply to non-registered users, who should have the means to prove their identity to the search engine provider*, for example, by registering for access to future data and/or *with a statement from their access provider about their use of a specific IP address in the past period about which access is requested.*[30]

IP addresses alone are not sufficient to reliably identify a user's searches for at least two reasons: multiple users and multiple locations. The first reason, the case of multiple users of the same IP address, is exemplified by a public computer, say at a library. There, many people use the same computer, and thus share the same IP address. A new cookie may be generated each time the web browser is re-opened after a prior user closes it, allowing the search engine to detect a possible change in user. However, without an actual username/password login, no actual identification is facilitated. The second reason that an IP address alone may be insufficient to track a user's queries, multiple locations for the same user, is exemplified by someone using the same laptop from different locations. A user may scatter his queries across multiple IP addresses, some of which he may own, some not. Again, without cookie information, and, in particular, an actual login, the user would not have access to his complete search history via IP address data alone.

IP addresses are still informative, however, as they can often be mapped to a small geographical region such as a county or zip code without requiring any non-public information. Here is an example from two free geolocation services (as of March 15, 2010):

[30] The Data Protection Working Party, *Opinion 1/2008 on Data Protection Issues Related to Search Engines*, WP148, 00737/EN (Apr. 4, 2008)1, 23 (emphasis added). Given that the search data are expected to remain confidential, it is not clear what purpose is served by allowing non-registered users access to prior searches.

Geolocation Service #1 - http://www.geobytes.com/IpLocator.htm?GetLocation
Geolocation Service #2 - http://www.melissadata.com/lookups/iplocation.asp

For a Stanford University IP Address [171.64.1.134]:

Service #1 came back with Palo Alto, CA, with a 99% certainty.
Service #2 came back with Stanford University.

For a U.C. Hastings IP Address [209.233.180.24]:

Service #1 came back with San Francisco, CA, with a 92% certainty.
Service #2 came back with San Francisco.

Search engine map information could also help associate IP addresses with physical locations to a reasonable degree of accuracy. Assume for the sake of argument that it is possible to identify at least a city or zip code from most IP addresses without having to get any third party information (e.g., from an ISP). While this might be useful for better searches, as discussed in Section II(A), *supra,* it creates problems under an assumption that an IP address personally identifies a user, as in the following hypothetical scenario.

Suppose a cafe owner wants to snoop into one of his customer's searches performed on the customer's laptop using the cafe's wireless access.[31] The owner contacts the search engine, validating that he owns the cafe's IP address, and reviews the search history originating from the cafe. Under this model, any valid IP address owner could similarly monitor the searches passing through the routers in their house, in their business, at their school, etc.[32]

[31] A typical setup is that a router at a business or in a home is given a public, dynamic IP address (that is, the external IP address might change from time to time). All the connections from internal computers to the external internet pass through the router, which assigns a local IP address to each internal computer. The only IP address seen by the outside world is the public one assigned to the router by the ISP, and the router keeps track of which incoming traffic should go to which local computer. Thus many computers can share the same external IP address simultaneously, making it impossible to know which local computer was involved in a given communication without knowing how the router handled internal addressing.

[32] This goes well beyond listening in on the network. A user is free to use encryption between their computer and a search engine (assuming for the sake of argument that the search engine allows encrypted searching). In this scenario, though, the

By assuming that an IP address safely identifies an individual, WP148 would open up more personal information than would be protected by deleting old IP addresses. IP addresses alone are simply not, in and of themselves, *reliably* personally identifying. Similarly, a user cookie is not reliably identifying.[33] From a privacy perspective, it is problematic to allow access to search histories based on that information alone. The right to review records should require at least a username and password – that is, be available only to registered users and their corresponding registered queries.

If IP addresses and user cookies for non-registered users are not reliably identifiable, then there is little need to delete them. For registered users, the search engine can directly associate the searches to the user, by the user's choice, and the IP addresses and cookies are not as determinant as the user's account information for identification purposes. In this case, though, there is no reason for the search engine to delete the data – the users should be able to do it themselves directly, if so desired, using their usernames and passwords.[34] In this manner, users can decide when to issue anonymous queries by first logging out, or deleting queries after-the-fact if they forgot to logout in advance. In either case, anonymization is unnecessary.

WP148, however, argues that if data can *ever* be used to identify someone, then they should be deleted.[35] This is a very high burden for any web site operator that logs IP addresses, as most do, and this requirement is not followed elsewhere. Clearly census data, health data, financial data, etc., are kept for decades, and much of that

eavesdropper goes to the search engine and simply identifies himself as the owner of the IP address and is then given full unencrypted access to all searches originating there.

[33] For example, public access computers may allow many people to reuse the same cookie for non-registered users. Similarly, a cookie that has never been associated with a particular person via a login is probably less identifying than the associated IP address.

[34] *See*, *e.g.*, http://www.nytimes.com/2009/03/11/technology/internet/11google.html ("Google will be the first major company to give users the ability to see and edit the information that it has compiled about their interests for the purposes of behavioral targeting.") (last visited Apr. 23, 2010).

[35] WP148 states that although "IP addresses in most cases are not directly identifiable by search engines, identification can be achieved by a third party." The Data Protection Working Party, *Opinion 1/2008 on Data Protection Issues Related to Search Engines*, WP148, 00737/EN (Apr. 4, 2008)1, 8. And then, quoting WP136, it goes on to state that unless a service provider can guarantee "with absolute certainty" that users cannot [ever] be identified by their IP address, "it will have to treat all IP information as personal data, to be on the safe side." *Id.*

contains personal identifiers.[36] As discussed in Section VI(A)(2), *infra*, describing the handling of U.S. medical data, the standard for identification there is that the data might *reasonably* be associated with an individual. Presumably the justification for the higher threshold for search query data (allowing for differences between the U.S. and the E.U.) might be a perception that they are more sensitive, or that the data are not needed, or that their usefulness does not warrant the perceived risk of keeping them. All of these reasons, however, are questionable, as discussed elsewhere in this article. If a reasonableness standard is sufficient to define personally identifying information, as a justification for the long-term retention of sensitive health information, it is difficult to justify the use of a much more stringent standard for search query data.

Ironically, though, another E.U. Directive, 2006/24/EC, mandates that European telecoms and ISPs keep records of all connections for up to two years, including phone numbers, locations, and/or IP addresses – even for unanswered calls.[37] Presumably these data are stored with the same level of protection as is required for health or financial data. If these private telecoms and ISPs are able to safely and confidentially store IP address information and conform to legal requirements, there is no reason to assume that the search engines would be any less capable of doing the same.

E. Anonymization is Ineffective

Anonymization is not only unnecessary, but also ineffective. There are broader issues related to internet privacy that society must solve in general. Without such solutions, search query anonymization is ineffectual; with such solutions, it is largely unhelpful.

It does not necessarily make sense to try to fix issues related to user privacy website by website or service by service, since many relate to the web in general. These include the following:

[36] *See* Section VI(A)(3), *infra* (discussing that financial data seems to be stolen much more frequently than health data. If the reason is due to the criminal desirability of the former compared to the latter, then search queries are not likely to be a primary target either.).

[37] The Directive mandates the retention of "data necessary to trace and identify the source of a communication." 2006/24, art. 5(1)(a), 2006 O.J. (L 105) (EC), *available at* http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2006:105:0054:0063:EN:PDF (last visited Apr. 23, 2010). This was brought about for national security reasons after the London subway terrorist bombings in 2005; *see*, *e.g.*, http://news.bbc.co.uk/2/hi/europe/4527840.stm (last visited Apr. 23, 2010). National security is discussed in Section II(B), *supra*.

> - cookie use, management, and transparency (best handled in the browser)
> - potential abuse of subpoenas, both 3rd party and governmental[38]
> - general logging of cookies and IP address information by all web sites

In addition, some search history is just inherently revealing, as demonstrated by the release of millions of "anonymized" AOL queries, following which the press was able to track down the identity of a user anyway.[39] Thus even anonymization does not fully solve the problem. Some may argue that, as a result, all query data should be immediately deleted (as impractical as that may be). However, a better argument may be that that (any) retained data should be given proper protection, rather than assuming that anonymization solves the problem, or even significantly helps. Since protection must be in place anyway in order to accommodate persistent personal data where deletion is not an option (e.g. email), anonymization is not necessarily the appropriate focus of privacy concerns.

As an example, in the section on "Some issues to be solved by industry" related to cookies, WP148 states the following:

> Persistent cookies containing a unique user ID are personal data and therefore subject to applicable data protection legislation. *The responsibility for their processing cannot be reduced to the responsibility of the user for taking or not taking certain precautions in his browser settings.* The search engine provider decides if a cookie is stored, what cookie is stored and for what purposes it is used. Finally, expiration dates of cookies set by some search engine providers seem to be excessive. For instance, several companies set cookies that expire after many years. When a cookie is used, an appropriate cookie lifetime should be defined both to allow an improved surfing experience and a limited cookie duration. *Especially in view of the default settings of browsers*, it is very important that users are fully informed about the use and effect of cookies. This information should be more prominent than simply being part of a search

[38] *See*, *e.g.*, Warshak v. U.S., 532 F.3d 521 (6th Cir. 2008) (en banc), further discussed in Section IV, *infra*.

[39] *See*, *e.g.*, Michael Barbaro & Tom Zellar Jr., *A Face Is Exposed for AOL Searcher No. 4417749*, N.Y. TIMES, Aug. 9, 2006, *available at* http://www.nytimes.com/2006/08/09/technology/09aol.html (last visited Apr. 23, 2010).

> engine's privacy policy, which may not be immediately apparent.[40]

These problems are endemic to the web, and search query anonymization will not solve them. Browser defaults need to be set appropriately so that they help with privacy immediately upon installation. For example, cookies should be deleted when the browser is closed. The fact that some cookies are persistent and unviewable, or that tracking-cookies monitor a user's activities and give the information to a third party for marketing purposes, seems much more of a concern to user privacy than having IP addresses or cookies stored in search query logs.[41] Security and user privacy would be greatly enhanced by transparent cookie use, a simple and prominent cookie management tool in the browser, and appropriate defaults for the novice user.

Another problem is the potential overuse or abuse of subpoenas to obtain search query data, as mentioned in Section II(B), *supra*. However, to reiterate here, if the problem is that we cannot trust the government, what would stop them from grabbing the information in transit and storing it anyway? If the problem is one of how the government or others might use the data, then we need a legal regime that restricts how the data can be used, including in civil cases. Anonymous access is not the same as anonymous storage, as exemplified by the use of the census data. The general problem of discovery of search query data is not so different from access to any other confidential information, whether it be email, library searches, medical history, etc. The problem would be better addressed by reining in allowable uses, rather than by attempts at anonymization. Again, anonymization or deletion will not solve the problem for any persistent data, and a viable legal framework needs to address the wider problem anyway.

Solving these problems would go a long way toward improving online privacy in general, and, in addressing them, allow for much greater search query privacy as well, without requiring data destruction.

[40] The Data Protection Working Party, *Opinion 1/2008 on Data Protection Issues Related to Search Engines*, Section 5.3, WP148, 00737/EN (Apr. 4, 2008) 1, 19–22 (emphasis added).

[41] *See, e.g.,* The Data Protection Working Party, *Opinion 1/2008 on Data Protection Issues Related to Search Engines*, Section 3, WP148, 00737/EN (Apr. 4, 2008), discussion regarding "flash cookies."

F. The Marketplace of Trust

Assuming that the data are useful and that anonymization is both unnecessary and ineffective, the proper solution to privacy may best be left, at least in part, to the market. As previously mentioned, in 2005 the U.S. Department of Justice sought a week's worth of anonymized user queries from AOL, Microsoft, Yahoo, and Google as part of its (unsuccessful) attempt to restore the Child Online Protection Act – ostensibly for the purpose of showing how many queries were related to pornography. Only Google fought the subpoena in court. In explaining their reasoning during an interview with ABC, Larry Page, Google's co-founder, stated that Google

> relies on having the trust of our users and using that information for that benefit. That's a very strong motivation for us. We're committed to that. If you start to mandate how products are designed, I think that's a really bad path to follow. I think instead we should have laws that protect the privacy of data, for example, from government requests and other kinds of requests.[42]

All three major search engines provide services such as email and chat. Arguably, the information contained in those conversations can be equally sensitive and private to the information found in search queries. And yet, no one argues that we should mandate that those conversations get deleted out of privacy concerns. Rather, we select companies based on our trust that they will keep such information private. Many factors may go into a user's selection of a particular search engine, such as the response time, the search result quality, the user interface, and the usefulness of the advertising. A reasonable consideration is also the level of trust that a user has regarding the search engine's handling of the user's privacy.

It seems rational that a company might seek to differentiate itself by showing that it cares about maintaining a user's trust. Although anonymization might be one valid method, using search history to improve query results while visibly fighting subpoenas is equally valid. Given that the search engines also have data that their users want them to store indefinitely, such as email, one could argue that the latter is a better indicator of an overall protection scheme than the former, since deletion says nothing about how a company will

[42] ABC News, *Bob Woodruff Reports From Inside the World of Google*, Jan. 20, 2006, http://abcnews.go.com/WNT/story?id=1526798 (last visited Apr. 23, 2010).

handle the data it keeps. A race to anonymization, whether it is 18 months, 6 months, or 2 days, might not be as convincing as developing best practices, including, say, public data protection audit ratings. Perhaps the differentiation is best left to the market to work out, under a framework of adequate data protection, rather than by imposing a particular solution, anonymization, on all the participants.

G. Collection Purpose

It is worth addressing the argument that anonymization is needed in order to prevent data from being used for a different purpose than that for which they were originally collected. The problem with this perspective is that it is not based on privacy. So long as the data are not being exposed to any additional parties, there is no privacy justification against repurposing.

Directive 95/46/EC highlights the importance of collecting, using, and retaining data for its "specified, explicit and legitimate" purpose.[43] WP148 summarizes the purpose restrictions as follows:

> In accordance with Article 6 of the Data Protection Directive [95/46/EC], personal data must be processed fairly and lawfully; they must be collected for specified, explicit and legitimate purposes and not be processed for purposes incompatible with the purposes for which they were originally collected. Moreover, the processed data must be adequate, relevant and not excessive in relation to the purposes for which they are collected and/or further processed. For any personal data processing to be lawful, it needs to satisfy one or more of the six grounds for legitimate processing set out in Article 7 of the said Directive.[44]

The aforementioned grounds include, for example, that the user has "unambiguously given his consent,"[45] or that the processing is "necessary for the purposes of the *legitimate interests* pursued by the controller."[46] This begs the question – what constitutes such legitimate interests? How broadly or narrowly are they interpreted, who should decide, and what is an appropriate doctrinal basis for answering these questions?

[43] Council Directive 95/46, art. 6(1)(b), 1995 O.J. (L 281) (EC).

[44] The Data Protection Working Party, *Opinion 1/2008 on Data Protection Issues Related to Search Engines*, Section 5, WP148, 00737/EN (Apr. 4, 2008)1, 15–22.

[45] Council Directive 95/46, art. 7(a), 1995 O.J. (L 281) (EC).

[46] Council Directive 95/46, art. 7(f), 1995 O.J. (L 281) (EC) (emphasis added).

In the process of analyzing the purpose of the data used by search engines, WP148 discusses some of the purposes given to them by several search engines in response to a survey.[47] These purposes include improving the service, securing the system, fraud prevention, law enforcement, etc. WP148 then discusses each in turn and comes to the following conclusion: "the Working Party does not see a basis for a retention period beyond 6 months."[48] As previously discussed, WP148 is used here as a comprehensive overview of many or most of the objections given to long-term or indefinite retention of non-anonymized search query data. As such, it is worth highlighting some of those objections in the context of a purpose-based argument in the hopes of extracting a possible basis for a legal doctrine that might allow for indefinite data retention. The general finding of the Working Party is summarized in the introduction to the section dealing with their analysis of the purpose of the non-anonymized search query data:

> Generally, search engine providers fail to provide a comprehensive overview of the different specified, explicit and legitimate purposes for which they process personal data. Firstly, some purposes, such as 'improvement of the service' or 'the offering of personalised advertising' are *too broadly defined* to offer an appropriate framework to judge the legitimacy of the purpose. Secondly, because many search engine providers mention many different purposes for the processing, it is not clear to what extent data are reprocessed for another purpose that is incompatible with the purpose for which they were originally collected.[49]

This is a case of a rejection of data retention due to a purpose being declared as "too broadly defined." Furthermore, it seems that mentioning different purposes for the same data, something that might be considered efficient from an engineering perspective, is somehow discouraged.

[47] The Data Protection Working Party, *Opinion 1/2008 on Data Protection Issues Related to Search Engines*, "Questionnaire for Search Engines on Privacy Policies," Annex 2, WP148, 00737/EN (Apr. 4, 2008) 1, 28–29.

[48] The Data Protection Working Party, *Opinion 1/2008 on Data Protection Issues Related to Search Engines*, "Search Engine Provider Obligation," Section 8, WP148, 00737/EN (Apr. 4, 2008) 1, 25.

[49] The Data Protection Working Party, *Opinion 1/2008 on Data Protection Issues Related to Search Engines*, "Analysis of Purposes and Grounds by the Working Party," Section 5.2, WP148, 00737/EN (Apr. 4, 2008) 1, 16–88 (emphasis added).

As WP148 deals with more details, more problems arise. For example, WP148 argues that more explicit consent can be construed from registered than non-registered users.[50] But if by consent we mean that users agree to use their information to answer a search query, then it is difficult to see how registration would matter. Assuming that the terms of service are the same for both registered and non-registered users, than the issuing of a query by its nature is a consent. Since, as discussed in Section II(D), *supra*, non-registered search queries arguably can not reliably identify a user, it is unclear what the problem is with retaining that data.

WP148 goes on to state that "it is the opinion of the Article 29 Working Party that search queries do not need to be attributable to identified individuals in order for them to be used to improve search services."[51] However, as discussed in Section II(A), *supra,* and Section II(I), *infra,* some type of identification is required to be able to track query reformulation across a query session. Furthermore, to the degree that IP addresses help in identifying a geographical region, that information might be very important for some forms of "local search" technology such as finding a nearby restaurant. The language of WP148 includes a dismissal of claims made by search engines that they need the data for accounting purposes.[52] However, accounting for costs among the billions of daily searches and billions of dollars requires cross-checking between various components of a large distributed system (e.g. checking that search data matches advertising data). It is complicated, and there is a lot of room for various forms of monitoring, validation, detection, fraud prevention, etc. It is not clear that the search engines are overstating the need for the data, or on what facts the Working Party bases their dismissal. In dismissing the need for the data for law enforcement, WP148 states the following:

> Moreover, large amounts of personal data in the hands of search engine providers may encourage law enforcement authorities and others to exercise their rights more often and

[50] The Data Protection Working Party, *Opinion 1/2008 on Data Protection Issues Related to Search Engines*, "-Consent-" subsection, Section 5.2, WP148, 00737/EN (Apr. 4, 2008) 1, 16–17.

[51] The Data Protection Working Party, *Opinion 1/2008 on Data Protection Issues Related to Search Engines*, "Legitimate Interest" subsection, Section 5.2, WP148, 00737/EN (Apr. 4, 2008) 1, 17–18.

[52] *Id*.

> more intensely which in turn might lead to loss of consumer confidence.[53]

This argument is, in effect, a government telling search engines to delete data because one cannot trust the government.[54] It goes on to tell search engines how best to succeed in the marketplace, which, as discussed in Section II(F), *supra*, might be something best left to the search engines to decide for themselves. The document goes on to state that "the Working Party considers that a reduced retention period will increase users' trust in the service and will thus constitute a significant competitive advantage."[55] Again, this seems to cross the boundary between a regulation based on protecting privacy into the realm of advice for competition in a marketplace. Moreover, if all search engines are forced to have the same short retention period, there would be no competitive advantage anyway.

The justification for anonymizing data is to protect user privacy, and any forced deletion of information should be grounded on that basis. When there is disagreement between the views of regulators and industry regarding the purposes for which the data are used, such discrepancies should be resolved within a framework of privacy protection. In that regard, one doctrine that might be valid is that a stated purpose be interpreted broadly so long as it does not extend the user's zone of access control. In other words, under an assumption of adequate data protection, if there is a reasonable interpretation of a stated purpose that is consistent with a proposed data use, then it should be allowed so long as doing so does not expose the data to a third party that would not otherwise have had access under a narrower interpretation.

If a company discovers an interesting new application from its existing data, the issue is whether that application results in sharing the data with anyone who would not have had access to personal information beforehand. For example, if a search engine is able to use search history to help with map lookups, and the data are shown

[53] The Data Protection Working Party, *Opinion 1/2008 on Data Protection Issues Related to Search Engines*, "Legitimate Interest" subsection, Section 5.2, WP148, 00737/EN (Apr. 4, 2008) 1, 17–18.

[54] Granted, though, that it might well be one government not trusting the actions of a different government. In any case, as pointed out *supra*, note 25, it is a somewhat odd argument that we should force the destruction of property not because we are concerned with abuse by the property owner, but with abuse by the government.

[55] The Data Protection Working Party, *Opinion 1/2008 on Data Protection Issues Related to Search Engines*, Section 5.3, WP148, 00737/EN (Apr. 4, 2008) 1, 19–22.

only to the registered user, who is given the option to opt in or out of the new application, no one is harmed – no extra data sharing has taken place.[56]

WP148 interprets restrictions of purpose as an argument to force data anonymization and deletion once the original narrowly interpreted purpose is completed. For example, use of personalized search query data to answer a query is fine, but using that same data to improve search results over several years apparently is not, let alone using the data in completely new ways and new applications. As with the other arguments, this perceived problem goes away under a model of adequate data protection, in which the data are stored intact and no more access is granted than had been allowed under the narrowly defined purpose. The cost of deletion in this case is not just quality of service, but innovations that could be as compelling and irreplaceable in the future as basic search has become today.

H. Whose Data Are They Anyway? – An American Constitutional Perspective

When a user searches for information, the query data gets logged. Although some of the data may pertain to the user, who is the proper "owner" of the log entry data? If governments impose data deletion as a privacy solution, should they have to compensate the search engines for the destruction of a valuable asset?

In the context of the user voluntarily giving the IP address and cookie, as is the case, the search engine has a proprietary right to record the information. Why does the user pass information such as IP address in the first place? Because the search engine has to know where to send the answer – to which machine. In fact, all the information that the search engine gets from the user may impact the search result content or format. For example, different browsers might have different capabilities and require different formats to display the results in the same way as other browsers. Similarly, a search cookie might help personalize search results. The search engine is simply recording the information sent by the user. That

[56] There might be an issue related to the value of the data. If someone gives information about themselves for one purpose, and a business uses it for an additional (albeit internal) purpose, the user might have a valid objection in terms of compensation. That is a contract issue, not a privacy concern. Another aspect of this is that of using the data in anonymized form in a new application. However, that is beyond the scope of this article since there are no privacy implications with such use, and pre-anonymization would allow the same use anyway.

information was given to them by the user in order to obtain the best search results.

It could be argued that such information is the rightful intellectual property of the search engine, in the same way that a medical clinic has the right to keep records of treatments given – in case of a claim of malpractice, to allow a doctor to review what she did in the past, to allow the clinic to audit procedural compliance, and, finally, to look at the aggregate of all the records to see which treatments worked best.

From a legal perspective, does the fact that information held by an entity is "about" a person give that person an inherent right to control the information? That is, is "aboutness" a property right, not just a privacy right?[57] One scenario is someone taking a photograph of another in a public place, or just writing down one's own thoughts about another, where the other person might not have any knowledge of the notes. Perhaps a more salient perspective is in situations where the information held is actually harmful to someone. Imagine the case of a person with a criminal record where a potential employer wants to know about it. A more common case is where a credit agency has (accurate) information about a person's bad credit. Perhaps the person would like that information deleted. We certainly cannot argue that we keep the information around to benefit him, but rather to protect lenders in an attempt to maintain an efficient financial system. We often allow someone to correct mistakes, say in credit information and criminal records. But we do not always allow even that – say in the case of non-public personal notes. Even for public writings, someone would only have cause to correct mistakes in cases of defamation, but not have the right to delete or alter personal observations by others that are factually correct (e.g., journalism). However we view it, there seem to be many instances in which "aboutness" does not translate to an ownership right, and data privacy is subject to a balancing of factors relative to which information we want others to be allowed to have about the object of the data (the "data subject" in E.U. terminology).

Assuming that search engines are proper owners of the data, search logs constitute a trade secret in that they are kept secret and derive value from their secrecy[58]. In *Ruckelshaus v. Monsanto Co.*,

[57] *See*, *e.g.*, Pamela Samuelson, *Privacy As Intellectual Property?*, 52 STAN. L. R. 1125, 1170–71 (2000).

[58] The Restatement of Torts defines a trade secret as "any formula, pattern, device or compilation of information which is used in one's business, and which gives him an

the U.S. Supreme Court held that the uncompensated taking of a trade secret by the EPA was unconstitutional:

> [Trade-secret] property right is protected by the Taking Clause of the Fifth Amendment. Despite their intangible nature, trade secrets have many of the characteristics of more traditional forms of property. Moreover, this Court has found other kinds of intangible interests to be property for purposes of the Clause.[59]

It is challenging to imagine another context in which the government has the right to force an entity to throw away its property, rather than regulate the use of that property, at least without the corresponding right of adequate compensation for the loss. The Court in *Ruckelshaus* found that:

> The question of what constitutes a "taking" is one with which this Court has wrestled on many occasions. It has never been the rule that *only* governmental acquisition or *destruction* of the property of an individual constitutes a taking . . . .[60]

The issue of a regulatory taking in this context is beyond the scope of this article, but the issues are interesting from an IP perspective.[61] However, allowing for data protection, rather than forced deletion, avoids this problem altogether.

opportunity to obtain an advantage over competitors who do not know or use it." Restatement (Second) of Torts § 757 cmt. b (1939).

[59] Ruckelshaus v. Monsanto Co., 467 U.S. 986, 987 (1984).

[60] *Id.* at 1004 (emphasis added). Based on the Court's wording, it seems that property destruction is assumed to be an obvious taking.

[61] As mentioned, one issue is whether the search engine is a rightful owner of their log data from a takings perspective. Another issue is whether forced deletion of the data would qualify as a *per se* taking. Under *Lucas v. South Carolina Coastal Council*, 505 U.S. 1003 (1992), it would qualify if all economic value was removed, and under *Loretto v. Teleprompter Manhattan CATV Corp.*, 458 U.S. 419 (1982), it would likely qualify under the concept of permanent dispossession of the property. *Penn Central Transportation Co. v. New York City*, 438 U.S. 104, 124 (1978) gives factors to consider in deciding the point at which regulatory restrictions become a taking, including the "economic impact of the regulation on the claimant and, particularly, the extent to which the regulation has interfered with distinct investment-backed expectations," as well as "the character of the governmental action." An important consideration is whether we would view the deleted data as independent, constituting 100% loss of property, or whether we would view the lost data as a component of the overall log data, constituting only a partial loss of the property's value. While a rational basis test is sufficient for a taking (Lingle v. Chevron U.S.A. Inc., 544 U.S. 528 (2005), Kelo v. City of New London, Connecticut, 545 U.S. 469, 487–88 (2005)), *Penn Central* implies that the point at which a regulation becomes a partial

I. Queries as User-Generated Content – A Moral Rights Perspective

Ironically, even the assumption that the user (co-)owns the data does not remove the problems caused by anonymization. At some point, complex query sessions become copyrightable user-generated content, and are thus subject to the moral rights of attribution and integrity.

Suppose that I investigate a research topic, spending an hour or more trying a few search queries, exploring the results, revising the queries, etc. For the purposes of illustration, I investigated the question of whether any of the Tuskegee Airmen were Jewish. I added some intentional misspellings and typos. It turned out to be a bit of a challenge to get an answer. For example, I could not find any data that gave the religious breakdown of the group. In the end, I came across a reference to one member who described himself as "a little Jewish," but found that by substituting "Judaism" for "Jewish" and then switching to searching through books rather than web pages:

Example Google query session (abridged): were any of the Tuskegee Airmen Jewish?[62]

tuskegie airmen
tuskegee airmen
<http://www.usnationalhistoricsite.com/>
<http://www.tuskegeeairmen.org/>
<http://en.wikipedia.org/wiki/Tuskegee_Airmen>
tuskegee airmen reliion
tuskegee airmen religion
tuskegee airmen jewish
<http://www.mlive.com/news/kalamazoo/index.ssf/2009/01/unexpected_opportunity_tuskege.html>

taking has to be viewed under a reasonableness standard. Under *Miller v. Schoene*, 276 U.S. 272 (1928), the government can force destruction of property without compensation. It is not clear, though, how that would be viewed if alternate solutions were available, or if there were not near 100% certainty of a nuisance or problem (in that case, treating rather than destroying fungus-infested trees; in this case, protecting rather than deleting data). Finally, another issue is whether or not a regulation that the data must be deleted, known in advance of the collection of the data, changes the anticipated value of the data going forward. In this case, the data must be collected, and all parties agree that it must be available for at least some number of months (for fraud detection, protection against DDOS attacks, etc.). Under *Palazzolo v. Rhode Island*, 533 U.S. 606 (2001), pre-existing regulations are not exempt from takings claims. The concept of a forward-looking governmental payment per deletion is consistent with the government paying rent for use of part of a private property, which is not uncommon.

[62] Each line represents either a query or a clicked-on URL. Indented lines are quotes from a given web page.

<http://www.blackjew.net/2009/01/barack-obama-invites-tuskegee-airman-to.html>

<http://www.jpost.com/servlet/Satellite?cid=1232292897063&pagename=JPost%2FJPArticle%2FShowFull>

<http://www.tuskegeeairmen.org/uploads/AirmanVisitsRockwoodCenter.pdf>

<http://ems.gmnews.com/news/2009/0225/bulletin_board/010.html>

tuskegee airmen biography jewish

<http://www.encyclopedia.com/doc/1E1-DavisBOJr.html>

<http://bajanreporter.blogspot.com/2009/01/barbadian-everywhere-one-of-tuskegee.html>

<http://blackhistorypages.net/pages/tuskair.php>

<http://www.southplainfield.lib.nj.us/homeworklinks/biography.htm>

<http://newpittsburghcourieronline.com/articlelive/articles/39948/1/Tuskegee-Airmen-facts/Page1.html>

tuskegee airmen religion|religious statistics

<http://www.josephgomer.com/>

<http://www.sandomenico.org/page.cfm?p=921>

<http://tuskegeeairmen.org/Tuskegee_Airmen_History.html>

<http://tuskegeeairmen.org/uploads/nameslist.pdf>

<http://www.mcall.com/news/local/election/la-me-tuskegee18-2009jan18,0,26561.story>

> “The Vietnam War in the 1960s was the first conflict to which the United States sent an integrated fighting force. [. . .] While Christianity was still the dominant religion among African-Americans within the military ranks, the institution had to accommodate its black soldiers’ other religions, including Hinduism, Islam, and Judaism.”

Encyclopedia of Religion and War - Google Book Search, http://books.google.com/books?id=WZdDbmxe_a4C&pg=PA68&dq=tuskegee+airmen%7Cairman+religion%7Cjew%7Cjewish&ei=XCizSeLZL4PKkQTtxbyqDg#PPA68,M1 p. 68 (last visited Mar. 7, 2009).

[moved to Google book search]

tuskegee airmen|airman jew|jewish|judaism

> "George Spencer Roberts was born in Fairmount, West Virginia, on September 24, 1918. He described himself as, 'Indian, black, Caucasian, a little Jewish.'"
>
> <http://books.google.com/books?ei=1CuzSYfSD4_AlQTS1Yy3Dg&id=LY9TAAAAMAAJ&dq=tuskegee+airmen|airman+jew|jewish|judaism&q=tuskegee+airmen|airman+jew|jewish|judaism&pgis=1#search_anchor>

Suppose that a search engine wanted to publish a book of their 1,000 most interesting query sessions. Ignoring for the moment the privacy issue, would the company need to ask the users for permission to do so strictly from a copyright perspective? If so, at the point at which copyright attaches to the query session, moral rights would presumably also attach – in particular, the rights of attribution (authorship) and integrity (non-destruction). If so, then anonymization or deletion initiated by the search engine should be precluded, since stripping a work of its authorship or destroying a work is contrary to these rights.

When does content become copyrightable? Generally, although laws vary between countries, there needs to be sufficient originality. While there is little debate that users own a copyright to their uploaded photographs, what about text they enter in a blog or on a friend's social networking page? While in general we may assume that email is copyrighted, information entered into forms is more complicated. For example, filling out one's name does not seem to carry with it a threshold level of originality. Similarly, a search query in isolation may be somewhat simple – "Tuskegee Airmen." But a series of queries taken as a whole, or in conjunction with a series of web page clicks (which may or may not be logged, depending on the search engine), might constitute a very unique form of expression. As we go from a simple "navigational" query such as "coca cola company" to a complex interaction involving query revisions, spell corrections, synonym expansions, image or map clicks, etc., the query session probably becomes sufficiently original to warrant copyright.[63]

[63] Further issues here include whether it matters that the content is generated in the context of functionality, though technical manuals and computer programs might fall under the same category and they are clearly copyrightable. Another issue is whether it matters if there was an intent by the user to ever display the "work". However, moral rights such as authorship would attach regardless of publication. Moreover, not only is there no requirement to register works for copyright to attach, but also if anyone else were to attempt to publish the work, the author's permission would likely still be required.

Moral rights are inherent to a work. In some countries, rights such as attribution are not alienable/negotiable. Even in the U.S., the Visual Artists Rights Act of 1990 ("VARA") provides for attribution and integrity rights of visual artwork, including the preclusion of destruction or mutilation of certain pieces.[64] International agreements under TRIPS (optionally) extend these rights to other copyrightable works.[65]

If we assume that moral rights attach to some query sessions, what might be the obligations of the search engine to protect them? For example, should it be required to maintain a copy, and, if so, for how long and under what conditions? We expect companies that store our email to keep it around, at least until the account is no longer active. But that expectation, and associated limitations, comes from a contractual agreement under the terms of service. In this case, the expectation derives from an inherent, non-negotiable right as a result of the creation of the content in the first place. In general, users do not maintain an independent copy of their search history. An additional interesting aspect of moral rights into the digital domain is that copies are identical. What does destruction mean in that context? Perhaps one way to view it would be to infer a notion of "last known copy". From this perspective, if someone believes or has reason to believe that they have the last digital copy of a work for which integrity attaches, they would be obliged to maintain the work.

Note that without the identifying information, it is impossible to string a query session together since there is no way to track which individual queries were issued from the same source.[66] They become a series of independent elements – like cutting up poetry into individual phrases. Even simple analyses such as discovering common term replacements or the contextual meaning of abbreviations is less accurate outside a query session model. In the context of moral rights, though, the loss of capability to view a query session as a single unit destroys the ability to reproduce the user-generated content. Thus forced anonymization is antithetical to a moral rights perspective.

[64] Visual Artists Rights Act of 1990 (VARA), 17 U.S.C. § 106A (2006).

[65] World Trade Organization, Overview: the TRIPS Agreement, http://www.wto.org/english/tratop_e/trips_e/intel2_e.htm (last visited Apr. 23, 2010).

[66] Query logs are generally recorded chronologically, possibly among thousands of search engine machines, which may then be merged into a central logging system. I do not assume that a user's queries, even within a single multi-query session, are handled by the same set of machines at the search engine.

Another issue arises pertaining to registered vs. non-registered users. Strictly speaking, moral rights attach to the creation of the work regardless of whether the search engine knows who the author is, as would be the case, say, with a book's author using a pseudonym (though presumably the publisher knows whom to pay). For the non-registered user, the IP address and search cookie associated with the query session are still valid. While they might not serve to inform the search engine about the identity of the author, they could in principle serve as protection against anyone else claiming authorship to the degree that the information can rule out a user who could not have had the IP address associated with the work. To the degree that a non-registered user's IP address and search cookie are not identifying, there is no reason to delete the information under the guise of protecting privacy. Thus, as before, a moral rights perspective would argue for keeping all the information to the degree that it aids in protecting the rights of attribution and integrity, and where that information serves no such purpose, the argument for deletion vanishes anyway. However, as with prior arguments, this apparent dichotomy goes away under a framework of protecting data rather than deleting it.

III. An Alternate Framework – HIPAA++

There are several factors to be considered in deciding on the minimal level of protection needed for search query data protection. One important issue is the fact that the E.U. and many in the U.S. are calling for anonymization within a few months of data collection, which would tend to push the data protection scheme toward a high level of security. Directive 95/46/EC allows for the use of "adequate" protection for non-E.U. members.[67] Thus, in principle, a framework

[67] Council Directive 95/46, art. 25(1), 1995 O.J. (L 281) (EC) requires that the non-E.U. country "ensures an adequate level of protection." Note, however, that the Data Protection Working Party, *Opinion 1/2008 on Data Protection Issues Related to Search Engines*, Section 4.1.2, WP148, 00737/EN (Apr. 4, 2008) claims jurisdiction over search data even in cases where the search query is passed to machines located outside the E.U. for processing. One justification given is that browsers perform some computation, and therefore some data processing goes on in the E.U. There are two potential flaws with this argument. First of all, whatever processing occurs on a personal computer happens prior to sharing the result of that processing with the search engine and it is not under the control of the search engine. Thus there is no relevance of such processing to the data given to the search engine. Second, computations done automatically and indirectly are no different than the processing that takes place in a video display control card. There are all kinds of processing that take place within a computer that the user does not directly initiate. Another jurisdictional justification is that the search engine front end machines,

in the U.S. that sufficiently safeguards search query data could be made to meet E.U. standards. This proposal uses health data as a model, since it has little financial desirability, can be very sensitive in nature, is not allowed to be given to third parties without an individual's authorization, and provides for legal sanctions for unauthorized disclosure.

The point of the framework, though, is to assign roles and distribute decisions. Certainly data deletion is one solution to the problem of protecting privacy by minimizing the possibility of disclosure.[68] However, assuming that there are other solutions, one approach is to let the government decide on the level of sensitivity and/or desirability of the data, and then mandate certain minimum security procedures for a given level of sensitivity. A company that wants to store the data can then decide if the cost of safeguarding the information at or above the minimum requirement is worth the perceived value of retaining it. This minimizes the role of the government from determining data usefulness and purpose beyond a broad area such as "medical use" or "information intermediation". Instead, the government's role is in assessing the sensitivity of the information and the level of threat or criminal interest (e.g., identity theft, espionage, terrorism, etc.). In this model, if we assume that medical information is more or less equally sensitive to search query information, and that their disclosure desirability are roughly equivalent, then there is little reason to impose a particular solution on one and not the other just because the government does not find the latter useful to keep around. Thus we do not want to set the level of protection unreasonably high in an attempt to force deletion, since

situated in Europe, generate a user cookie, and that that qualifies as personal data processing. These cookies, however, are sent to both registered and non-registered users, and, as has been discussed *supra*, Section II(B), are not in and of themselves personally identifying. It is difficult to see how generating a cookie is any more relevant to jurisdiction than is underlying network processing. The personalization happens only by associating the cookie with a registered user, which does not happen in the process of simply generating a cookie. It would be like saying that generating a timestamp justifies jurisdiction. For a different perspective on jurisdictional approaches, see the Canadian framework: Office of the Privacy Commissioner of Canada, Guidelines for Processing Personal Data Across Borders, http://www.privcom.gc.ca/information/guide/2009/gl_dab_090127_e.asp (last visited Apr. 23, 2010).

[68] Remember that the data have to be stored for a few months for several reasons such as click fraud and DDOS detection, and thus disclosure is possible during that window. Furthermore, as previously mentioned, there is the possibility of capturing the data in transit or making illicit copies prior to deletion. The only reliable way to protect queries is not to make any in the first place. Beyond that, there are no 100% solutions.

that solution would not accommodate persistent sensitive information in general, such as medical data.

Legal Framework Suggestion – HIPAA++:

- Search query data (non-anonymized) should be held to no higher protection standard than medical data – that is, the protection standard should be determined by the sensitivity and theft/disclosure desirability, not by the use or (narrow) purpose.
- No release of data to third parties without specific opt-in authorization
- Anonymized access, not anonymized storage.
- Presumption of non-identification: IP addresses should not be considered as reliably identifiable in and of themselves. Similarly for non-registered cookies.
- Discoverability: data limited to criminal cases with a showing of probable cause.[69]
- Mandatory employee training, compliance officers, and data protection audits.
- Agency and individual civil claims, punitive damages, criminal liability.

Technical Framework Suggestion (e.g. HIPAA/NIST):

- Segregate and encrypt personally identifying data.
- Non-personal data maintains pointers to personal data.
- Access via an API with appropriate access control restrictions.
- Full monitoring of access to personal data.

This skeletal suggestion is simply to give an example of a framework that provides a reasonably high level of data protection. One way or another, reasonable protection is possible without anonymization.

IV. Shooting Ourselves in the Foot

A. An Interesting Case Study

There has been ongoing litigation regarding the Stored Communications Act ("SCA")[70] allowance of government *ex parte* searches of email without probable cause; if not for a finding of lack of ripeness, it very well might have been found unconstitutional in

[69] If excluding the use of these data for civil purposes seems extreme, keep in mind that doing so is less drastic than actual deletion.

[70] Discussed in Section VI(A)(4), *infra*.

Warshak v. U.S. by the Sixth Circuit (as it was, both facially and as applied, in the 3-judge panel prior to the en banc hearing).[71] In that case, the government declined to notify the suspect about the email searches for one year, well past the statutorily-required 90-day maximum. In throwing out the constitutional challenge, the court presented an interesting perspective on a user's right to expect privacy and the role of service providers:

> Think of just one of these moving parts-the variety of internet-service agreements and the differing expectations of privacy that come with them. An agreement might say that a service provider will "not . . . read or disclose subscribers' e-mail to anyone except authorized users." *United States v. Maxwell*, 45 M.J. 406, 417 (C.A.A.F.1996) (describing testimony about AOL's then-existing policy). An agreement might say that a service provider "will not intentionally monitor or disclose any private email message" but that it "reserve[s] the right" to do so in some cases. *See* Privacy Statement for Juno Members, http://www.juno.com/legal/privacy.html (last visited July 7, 2008). An agreement might say that a service provider "may or may not pre-screen Content, but . . . shall have the right (but not the obligation) in [its] sole discretion to pre-screen, refuse or move any Content that is available via the Service"-as indeed Warshak's Yahoo! account did. JA 89, 163 n. 3. An agreement might say that e-mails will be provided to the government on request-as indeed the same Yahoo! account did. An agreement might say that other individuals, besides the recipient of the e-mail, will have access to it and will be entitled to use the information in it. *See, e.g.,* JA 208 (explaining that Gmail, a service provided by Google, gives users "an enormous amount of storage capacity . . . in exchange for . . . terms of service which say that Google is allowed . . . [to] take a look at the content of [users'] e-mail and . . . target advertising at [users] accordingly"). Or an agreement might say that the user has no expectation of privacy in any of her communications. *See, e.g.,* JA 207 (government counsel explaining that "every day when we log into our e-mail account, we agree that we have no expectation of privacy in the account").[72]

The dissent made a compelling rebuttal:

> The majority adequately recites the facts, but conveniently leaves out what I believe to be an essential element of the case.

[71] *Warshak v. U.S.,*, 532 F.3d at 521.

[72] *Warshak v. U.S.,* 532 F.3d 521 at 527.

> As the majority correctly states, § 2703(d) allows a court to issue an order based on less than probable cause, allowing the government to search a suspect's email communications stored with an electronic service provider for more than 180 days. Typically, in order to effect such a search, the Stored Communications Act requires the government to notify the suspect of the search. However, § 2703(b)(1)(B) allows the court to grant the government a 90-day delay of the notice if notification would result in "(A) endangering the life or physical safety of an individual; (B) flight from prosecution; (C) destruction of or tampering with evidence; (D) intimidation of potential witnesses; or (E) otherwise seriously jeopardizing an investigation or unduly delaying a trial." 18 U.S.C. § 2705(a)(2), (b). What the majority leaves out is the fact that while the government was initially granted a 90-day delay before being required to notify Warshak of its searches of his email accounts, when the 90 days expired, the government ignored the statute and failed to notify Warshak of its searches. Over a year went by before Warshak became aware that his emails had been searched. While members of this Court may argue over whether or not the delayed notification section of the Stored Communications Act is constitutional, it is uncontroverted that the government violated the law by failing to notify Warshak 90 days after searching his emails.
>
> The fact that the government was unable to abide by an arguably unconstitutional provision of the Stored Communications Act informs any analysis of Warshak's motion for preliminary injunction. Not only is Warshak alleging that the delayed notification provision of the act is unconstitutional, but he is also alleging that the government cannot be trusted to abide by the actual requirements of that law as written.[73]

If the goal of the search engines is to retain data, then they certainly are not helping to make that happen with terms of service that do not guarantee an expectation of privacy to their users. Beyond what may be provided as a result of market forces and public pressure, the E.U. is demanding adequate privacy protection. If they cannot get it through adequate safeguards, perhaps they are right in demanding data destruction. It seems that in this regard, the search engines are working against their own interests.

The government is also working against its own interests. By legislating exceptions to the warrant requirement, by ignoring protections that exist during investigations and prosecutions, and by

[73] *Warshak v. U.S.*, 532 F.3d 521 534–5 (Martin, J., dissenting).

refusing to declare such activities unconstitutional, the government may well force deletion of data that they would otherwise be able to use down the road. The result of cases like *Wayshak* may be a hesitancy to leave data with service providers. And beyond that, it shows to the world that the U.S. does not take privacy seriously – leaving data destruction as the only viable solution to privacy protection. This is counterproductive to law enforcement and national security if the data are useful for those purposes. Similarly, users who do not adamantly demand privacy in their terms of service, transparent cookie usage, and clear safeguards, give courts reason to say that there is no expectation of privacy. That is contrary to their interest. Even the potential for jurisdictional over-reaching[74], if true, is potentially counter-productive – if a country with strong privacy protection can over-reach, so can a country with weak privacy protection.

A framework of adequate privacy protection is a reasonable alternative to data destruction. Anything less invites purging of data that is invaluable to all stakeholders.

V. Conclusion

Taken individually, each argument presented here has various strengths and weaknesses. However, two points can probably be made safely. First, there are clearly issues other than privacy that are impacted by anonymization, whether forced or not. These include the loss of potentially useful data, the benefits of repurposing, data ownership, and more. Justifying anonymization solely in the name of privacy ignores the trade-offs being made for a less than ideal solution. Second, when taken as a whole, these arguments tend to reinforce each other such that their combination is more convincing. They serve to question the reasonableness of anonymization as a privacy solution, at least to the degree that other methods should be examined more closely.

Protecting privacy is important, necessary, and possible. Hopefully it can be done with full recognition of the value of search query data. A balanced approach would minimize unauthorized disclosure of sensitive data, allow for the myriad benefits of long-term storage, and serve each stakeholder's best interests.

[74] European Union, *supra* note 67.

VI. Appendix – Data Protection Schemes

In contrast to search queries, where identifying or partially identifying information is apparently deemed to be unimportant enough that its deletion is acceptable, the perceived necessity of long term retention of many other types of data is different. For example, medical data are taken to be inherently useful in order "to provide and promote high quality health care and to protect the public's health and well being."[75] Similarly, "accurate credit" information is needed to assure "fair and accurate credit reporting," which is "essential to the continued functioning of the banking system."[76] As will be discussed below, the basis for this data retention includes benefits to the collective, even if at times it is detrimental to the individual. We review here some of the mechanisms used to protect both data integrity and confidentiality in other arenas, namely, in the U.S.: census data, medical data, financial and credit data, and electronic communications and storage data. We then take a brief look at the E.U. framework for data protection, leading to the introduction of the aforementioned WP148 document on search query privacy.

A. Data Protection in the U.S.

Data protection in the U.S. is comprised of a piecemeal collection of various acts, regulations, and standards. One difference between these approaches is whether the information is available outside the controlling authority at all (e.g., no for census data, yes for medical and financial data), and, if so, whether a search warrant is needed to access it (e.g., yes for electronic communications). Another difference is whether the data are available to third parties involved in the initial activity for which the data were collected (e.g., medical data available to health insurance companies for purposes of payment), or available to sell to third parties without permission of the individual (e.g., financial data where no "opt out" notice has been signed). Some statutes specifically require training regarding confidentiality (e.g., medical data), and some provide for criminal penalties for giving data to unauthorized persons (e.g., medical data, census data).

[75] U.S. Department of Health and Human Services, Office for Civil Rights, *Summary of the HIPAA Privacy Rule*, http://www.hhs.gov/ocr/privacy/hipaa/understanding/summary/privacysummary.pdf (last visited Apr. 23, 2010).

[76] Fair Credit Reporting Act, 15 U.S.C. § 1681 (a)(1) (2006).

The purpose of this brief introduction to U.S. data protection is to build up intuition for the relationship between the perceived sensitivity of the information and the level of protection required. Without stating it directly, the patchwork of U.S. data protection legislation is somewhat consistent with the general provision described in the E.U. Directive 95/46/EC for the relationship between data sensitivity and its corresponding requisite security:

> Having regard to the state of the art and the cost of their implementation, such measures shall ensure a level of security appropriate to the risks represented by the processing and the nature of the data to be protected.[77]

This relationship between the level of sensitivity and the degree of security is discussed in more detail in Section III. Part of the discussion related to making security commensurate with sensitivity is whether or not there should be consideration given to the perceived usefulness or purpose of the data, and, if so, who should decide how useful the data might be, and to which purpose it may contribute. Thus, as we review some of the data protection schemes below, it is worth considering the implied value of the data, and by whom that value determination is made.

1. Census Data

Under Sections 9 and 214 of Title 13 of the U.S. Code, census data are held to a rigorous standard of protection, including criminal liability of fines and imprisonment. The Census Bureau maintains a Data Protection website that describes the legal, procedural, and statistical safeguards used to maintain confidentiality.[78] According to their online catalog, they retain data going back to the colonial times of the 1790s.[79] Under Section 221, answering census surveys is mandatory. According to the Census Bureau's FAQ (Frequently Asked Questions), census data seem to be stored in their entirety but

[77] Council Directive 95/46, art. 17(1), 1995 O.J. (L 281) (EC).

[78] U.S. Census Bureau, Data Protection and Privacy Policy, http://www.census.gov/privacy/data_protection/how_we_protect_your_information.html (last visited Apr. 23, 2010).

[79] U.S. Census Bureau, Census Product Catalog, Reference, http://www.census.gov/mp/www/cat/reference/index.html (last visited Apr. 23, 2010).

released only in aggregate, and the data are not allowed to be released to other government agencies or third parties in raw form.[80]

2. *Medical Data*

Medical information stored with health care providers, health plans, and health care "clearinghouses," as well as their relevant business associates, is often intended to be kept at least for the life of the patient, but under a relatively high level of privacy protection.[81] Congress passed the *Health Insurance Portability and Accountability Act* ("HIPAA") in 1996:

> A major goal of the Privacy Rule is to assure that individuals' health information is properly protected while allowing the flow of health information needed to provide and promote high quality health care and to protect the public's health and well being. *The Rule strikes a balance that permits important uses of information, while protecting the privacy of people who seek care and healing.* Given that the health care marketplace is diverse, the Rule is designed to be flexible and comprehensive to cover the *variety of uses* and disclosures that need to be addressed.[82]

Thus at least in the case of medical data, the government recognizes not only the need to balance use with privacy protection, but also that the data has many uses that need to be accommodated. The rules are designed to prevent data abuse while allowing reasonable use:

> Under the patchwork of laws existing prior to adoption of HIPAA and the Privacy Rule, personal health information could be distributed—without either notice or authorization—for reasons that had nothing to do with a patient's medical treatment or health care reimbursement. For example, unless otherwise forbidden by State or local law, without the Privacy Rule patient information held by a health plan could, without the patient's permission, be passed on to a lender who could then deny the patient's application for a home mortgage or a credit card, or to an employer who could use it in personnel

[80] U.S. Census Bureau, Question and Answer Center, http://ask.census.gov/cgi-bin/askcensus.cfg/php/enduser/std_adp.php?p_faqid=781 (last visited Apr. 23, 2010).

[81] 45 C.F.R. § 160.103 (2009).

[82] U.S. Department of Health and Human Services, Office for Civil Rights, *Summary of the HIPAA Privacy Rule*, http://www.hhs.gov/ocr/privacy/hipaa/understanding/summary/privacysummary.pdf (last visited Mar. 10, 2009) (emphasis added).

> decisions. The Privacy Rule establishes a Federal floor of safeguards to protect the confidentiality of medical information.[83]

The rules accommodate use for medical reasons while blocking non-medical use. This is a broad interpretation of the purpose for which the data are collected, and is not limited to direct patient care. For example, nothing in the rules forbids a health provider from looking over data in the aggregate to investigate treatment success.

The protected information is described as follows:

> The Privacy Rule protects all "individually identifiable health information" held or transmitted by a covered entity or its business associate, in any form or media, whether electronic, paper, or oral. The Privacy Rule calls this information "protected health information (PHI)." 45 C.F.R. § 160.103.
>
> "Individually identifiable health information" is information, including demographic data, that relates to:
>
> - the individual's past, present or future physical or mental health or condition,
> - the provision of health care to the individual, or
> - the past, present, or future payment for the provision of health care to the individual,
>
> and that identifies the individual or for which there is a *reasonable basis* to believe it can be used to identify the individual. 45 C.F.R. § 160.103.[84]

Note that in this case, there must be at least a reasonable basis in believing that the information can be tied to an individual. This stands in contrast to search query data, where some have argued that *any possibility* of associating information with an individual is sufficient to label it as identifying.[85]

Furthermore, anonymized ("de-identified") medical information under HIPAA is allowed unrestricted use.[86] Such data are defined as

[83] U.S. Department of Health and Human Services, Why is the HIPAA Privacy Rule needed?, http://www.hhs.gov/ocr/privacy/hipaa/faq/about/188.html (last visited Apr. 23, 2010).

[84] *Id*. (emphasis added).

[85] *See* Section II(D), *supra*.

[86] 45 C.F.R. § 164.502 (2009).

"information that does not identify an individual and with respect to which there is no reasonable basis to believe that the information can be used to identify an individual."[87] For example, zip code information must be modified so as not to contain groupings of less than 20,000 people.[88] HIPAA details the allowable uses of protected information both with and without the individual's consent.[89] Furthermore, the Act spells out requirements for data protection, including company contact personnel responsible for implementing data security and receiving and acting on an individual's request to see and/or modify his own information.[90] Companies are required to mitigate any data security breach, and must "reasonably safeguard" relevant data.[91] The safeguards are more formally described by both HHS and NIST.[92] Finally, HIPAA provides for both civil and criminal penalties for failure to implement appropriate data security.[93] Nothing in the Act was found to speak of data deletion, but only of anonymization of released data prior to unrestricted use.

3. Financial Data

a. Financial Services Modernization Act

One part of the Financial Services Modernization Act ("FSMA"), passed by Congress in 1999, involves the privacy of financial data. The policy underlying this data protection is that "each financial institution has an affirmative and continuing obligation to respect the privacy of its customers and to protect the security and confidentiality of those customers' nonpublic personal information."[94] Each financial institution is obliged to implement safeguards for their customer data:

> In furtherance of the policy in subsection (a) of this section, each agency or authority described in section 6805(a) of this title [15] shall establish appropriate standards for the financial

[87] 45 C.F.R. § 164.514(a) (2009).

[88] 45 C.F.R. § 164.514(b)(2)(i)(B) (2009).

[89] 45 C.F.R. § 164.502 (2009).

[90] 45 C.F.R. § 164.530(a), (b) (2009). *See also* 45 C.F.R. §164.520(b) (2009).

[91] 45 C.F.R. § 164.530(c), (f) (2009).

[92] U.S. Department of Health & Human Services, Health Information Privacy, http://www.dhhs.gov/ocr/privacy/hipaa/administrative/securityrule/securityruleguidance.html (last visited Apr. 22, 2010).

[93] 42 U.S.C. § 1320d-5, -6 (2006).

[94] 15 U.S.C. § 6801(a) (2006).

> institutions subject to their jurisdiction relating to administrative, technical, and physical safeguards —
>
> > (1) to insure the security and confidentiality of customer records and information;
> > (2) to protect against any anticipated threats or hazards to the security or integrity of such records; and
> > (3) to protect against unauthorized access to or use of such records or information which could result in substantial harm or inconvenience to any customer.[95]

Enforcement of these provisions are also described in the Act, and does not include criminal liability.[96] Although most of the "nonpublic personal information" may be shared with (i.e. sold to) third parties if a customer does not "opt out", account numbers are not allowed to be shared for marketing purposes regardless.[97]

b. Fair Credit Reporting Act

In addition to the FSMA protection of financial data, the Fair Credit Reporting Act ("FCRA") regulates credit information and the nationwide credit reporting agencies. The goal of the Act is to assure fair and accurate consumer credit reporting, with "respect for the consumer's right to privacy" using "reasonable procedures" performed "in a manner which is fair and equitable to the consumer, with regard to the confidentiality, accuracy, relevancy, and proper utilization of such information."[98] For example, all consumers are allowed to obtain a free annual credit report from each of the three major credit reporting agencies.[99]

The FCRA limits how credit information may be used (e.g., legitimate credit checks, court orders, etc.).[100] It also regulates which information is allowed or required to be included in a credit report (e.g., disallowing bankruptcy and litigation information that is too old).[101] Enforcement of these provisions includes civil liability,

[95] 15 U.S.C. § 6801(b) (2006).

[96] 15 U.S.C. § 6805 (2006).

[97] 15 U.S.C. § 6802(b), (d) (2006).

[98] 15 U.S.C. § 1681 (2006).

[99] *See, e.g.*, Federal Trade Commission, Your Rights: Credit Reporting, http://www.ftc.gov/bcp/menus/consumer/credit/rights.shtm (last visited Apr. 21, 2010) and Annual Credit Report, http://www.annualcreditreport.com (last visited Apr. 21, 2010).

[100] *See* 15 U.S.C. § 1681b(a) (2006).

[101] 15 U.S.C. § 1681c (2006).

including attorney fees, and punitive damages for willful violations.[102] Knowing or willful release of a consumer's information to an unauthorized person can also result in criminal liability.[103] In addition to direct consumer litigation, civil action can also be brought by several federal agencies.[104]

It is somewhat difficult to reconcile the high level of protection and enforcement of credit information, which includes punitive damages, criminal liability, and agency enforcement, with the relatively weaker enforcement provisions and "opt out" mechanism of financial information. Identity theft is commonly in the news and one of the FTC's highest priorities, accounting for hundreds of thousands of complaints annually in the U.S. alone.[105] Given the cost of unauthorized financial information disclosure and the desirability of the data to criminals, perhaps such desirability should be considered a more important factor in determining the level of data security required than usefulness or purpose – this approach is more consistent with threat analysis methodology.[106]

4. Communications Data

a. Electronic Communications Privacy Act; Stored Communications Act

The Electronic Communications Privacy Act ("ECPA"), 18 U.S.C. § 2510 et seq., protects communication of individuals from government surveillance undertaken without a court order, from being accessed by third parties lacking legitimate authorization to access the messages, in addition to protection from access by message carriers, such as ISPs.[107] As part of the ECPA, the Stored Communications Act ("SCA"), 18 U.S.C. § 2701 et seq., "addresses

[102] 15 U.S.C. § 1681n, o (2006).

[103] 15 U.S.C. § 1681r (2006).

[104] 15 U.S.C. § 1681s (2006).

[105] *See, e.g.*, Federal Trade Commission, *Fiscal Year 2009 Congressional Budget Justification*, at 1, *available at* http://www.ftc.gov/ftc/oed/fmo/budgetsummary09.pdf ("[I]dentity theft remains on top of the FTC's list of consumer complaints, accounting for more than 34 percent of the 728,765 complaints (not including Do Not Call Registry complaints) filed in FY 2007.").

[106] One of the major steps in analyzing a system's security is, understandably, identifying potential threats. *See, e.g.*, J.D. Meier, Alex Mackman, and Blaine Wastell, Threat Modeling Web Applications, May 2005, http://msdn.microsoft.com/en-us/library/ms978516(v=MSDN.10).aspx.; Syed Naqvi and Michel Riguidel, *Threat Model for Grid Security Services*, *in* LECTURE NOTES IN COMPUTER SCIENCE 1048.

[107] *See* 18 U.S.C. §§ 2511, 2516, 2517 (2006).

voluntary and compelled disclosure of 'stored wire and electronic communications and transactional records' held by third-party internet service providers ("ISPs")."[108] As one law review article described it, "courts, legislators, and even legal scholars have had a very hard time understanding the method behind the madness of the SCA. The statute is dense and confusing, and that confusion has made it difficult for legislators to legislate in the field, reporters to report about it, and scholars to write scholarship in this very important area."[109] Let this article be no exception. However, it's worth noting that the SCA is oriented toward communication in "temporary, intermediate storage" or stored "by an electronic communication service for purposes of backup protection."[110] While the SCA may include various protections against government access of a user's email stored at a search engine (e.g., gmail, hotmail, yahoo mail, etc.), it probably does not extend to search query logs.

The ECPA is an extension of the Federal Wiretapping Act, protecting communications in transit, though "the Circuits that have addressed the issue have agreed that the definition of 'intercept' 'encompasses only acquisitions contemporaneous with transmission.' *United States v. Steiger*, 318 F.3d 1039, 1047 (11th Cir. 2003)."[111] The SCA protects stored communications such as email. Some divergence of opinion exists between the circuit courts regarding the temporary storage of email while in transit as to whether it can be considered simultaneously in transit and in storage. See *U.S. v. Councilman*,[112] and *Bailey*[113] for an explanation of the ambiguity of the statutes and the differences between the circuits. One interesting issue raised in *Bailey* is the following:

> However, as a point of clarification, Stored Communications Act protection does not extend to emails and messages stored only on Plaintiff's personal computer. *In re Doubleclick Inc.*, 154 F.Supp.2d 497, 511 (S.D.N.Y. 2001) ("the cookies' residence on plaintiffs' computers does not fall into §

[108] Wikipedia, Stored Communications Act, http://en.wikipedia.org/wiki/Stored_Communications_Act. *See also* 18 U.S.C. §§ 2702, 2703 (2006).

[109] Orin S. Kerr, *A User's Guide to the Stored Communications Act, and a Legislator's Guide to Amending It*, 72 GEO. WASH. L. REV. 1208, 1208 (2004).

[110] Hilderman v. Enea TekSci, Inc., 551 F. Supp. 2d 1183, 1204–05 (S.D. Cal. 2008) (quoting 18 U.S.C. § 2510(17)).

[111] Bailey v. Bailey, No. 07-11672, 2008 WL 324156, at *4 (E.D. Mich. Feb. 6, 2008).

[112] U.S. v. Councilman, 418 F.3d 67 (1st Cir. 2005).

[113] *Bailey*, 2008 WL 324156, at *6.

> 2510(17)(B) because plaintiffs are not 'electronic communication service' providers.").[114]

Thus is seems that cookies are included as part of the communication regulations, though not the ones on a user's computer. That is, assuming that search engines might be considered an "electronic communication service" provider, it might be considered illegal for them to share such information with third parties.

That being said, the ECPA and SCA require search warrants for most types of governmental access to electronic communication such as email, with exceptions under section 2703(d) for ongoing investigations.[115] In terms of the overall discussion of U.S. data protection, these statutes are representative of some of the highest levels of restrictions, where applicable, in requiring warrants for some types of data access. Such restrictions are one way to protect against intrusion into sensitive information such as search query data.

B. Data Protection in the E.U.

As mentioned in the Introduction, data protection in the E.U. is structured around Directive 95/46/EC, the so-called "Data Protection" directive.[116] While details of the directive are discussed elsewhere in this article, the overall framework is quite general.[117] Other related directives include Directive 2002/58/EC, regarding "the processing of personal data and the protection of privacy in the electronic communications sector"[118] and Directive 2006/24/EC "on the retention of data generated or processed in connection with the provision of publicly available electronic communications services or of public communications networks."[119] These last two directives regulate electronic communications, similar in scope to the ECPA and SCA in the U.S.

[114] *Id*.

[115] 18 U.S.C.A. § 2703 (Supp. 2009). See further discussion of the warrant issue in Section IV, *supra*, quoting from *Warshak v. U.S.*, 532 F.3d at 527, 534–35 (en banc).

[116] For the E.U.'s official privacy directives, *see* Justice and Home Affairs: Data Protection, http://ec.europa.eu/justice_home/fsj/privacy/index_en.htm (last visited Apr. 21, 2010).

[117] *See*, *e.g.*, "Data Protection in the European Union," http://ec.europa.eu/justice_home/fsj/privacy/docs/guide/guide-ukingdom_en.pdf (last visited Apr. 21, 2010).

[118] Council Directive 2002/58, 2002 O.J. (L 201) 37 (EC).

[119] Council Directive 2006/24, 2006 O.J. (L 105) 54 (EC).

1. European Directive 95/46/EC and the Article 29 Working Party Report WP148

The Data Protection Directive introduces several concepts relevant to the discussion of data protection and information privacy. For example, a "data subject" is the person described or identified by the data, such as in the case of someone's social security number. A "data controller" is the person who has control over the physical data, such as the owner of a database. A "data processor" is any entity that is involved with processing the data, which may be, for example, a third party hired by the data controller. Finally, the "data purpose" constitutes the reason that the controller collected the data about the data subject.

As mentioned in the introduction, Directive 95/46/EC authorized the creation of the Article 29 Working Party advisory committee. This committee consists of representatives of E.U. member countries and issues various opinions and reports on an as-needed basis. One such particularly relevant report, as discussed, is WP148, from 2008, which presents a comprehensive overview for the case of search query anonymization. The report concludes that search query data should be anonymized within six months.