

Search Query Privacy: The Problem of Anonymization

Ron A. Dolin, J.D., Ph.D.*

Research performed at University of California, Hastings College of the Law
Ron.A.Dolin@gmail.com

Abstract

This paper presents several arguments that the deletion or anonymization of search query data is problematic. A more balanced 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.

Introduction

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 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 (e.g. <http://www.google.com/intl/en/privacy.html>), search queries and other user-generated content have been the subject of governmental, private, and international discovery.¹ As a result, many

countries have initiated 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 it is not available for later discovery or other uses.² Privacy advocates are optimistic in pushing a legislative agenda in this direction, and Representative Edward J. Markey, ex-chairman of the House Subcommittee on Telecommunications and the Internet, has suggested adopting similar policies.³

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 paper sets out several issues suggesting that the push toward anonymization and deletion is ill-conceived. The paper also suggests borrowing from, and adding to, other data protection schemes, such as that used for health records, that serve to maintain

queries.html, and <http://www.paed.uscourts.gov/documents/opinions/07D0346P.pdf>. In another well-known case, in 2008 Viacom sought, and 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.*, <http://www.nytimes.com/2008/07/04/technology/04youtube.html>. 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.*, <http://www.sfgate.com/cgi-bin/article.cgi?f=/c/a/2007/11/07/MN2NT7C99.DTL>.

- 2 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 cannot be traced back to an individual machine.
- 3 MIGUEL HELFT, Yahoo Limits Retention of Search Data, *New York Times*, <http://www.nytimes.com/2008/12/18/technology/internet/18yahoo.html> (Dec. 17, 2008).

Copyright © 2010 Ron A. Dolin. All rights reserved.

* 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. A more detailed version of this paper is to appear in the Summer, 2010 issue of the *Hastings Science & Technology Law Journal* (Vol. 2, Issue 2).

- 1 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 leading four search engines in order to establish the percentage of searches related to pornography – only Google fought the subpoena. *See, e.g.*, http://en.wikipedia.org/wiki/Child_Online_Protection_Act, <http://googleblog.blogspot.com/2006/03/judge-tells-doj-no-on-search->

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 forced data destruction.

Anonymization Problems⁴

Usefulness

Enforcing privacy through anonymization comes at the cost of any benefit derived from the use of non-anonymized data. In a nutshell, it is difficult to overstate the vast potential of uses of search query information, limited only by one's imagination. The data are not only valuable to the user, 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, to defend 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.).⁵ IP addresses and cookies are important in all of these.

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).⁶ The work is accurate enough to have warranted publication in the well-known science journal *Nature*.⁷ As discussed below, IP ad-

resses 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 saves lives. This tool was developed by looking over 5 years worth of data.

Another example, perhaps just for curiosity, is what people are searching for in your local area.⁸ However, imagine looking for a 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 probably 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.⁹ 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 are 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.

Exceptions to Deletion

The data are useful to many stakeholders. If we follow the path of the E.U., we will be left with a serious dilemma. On the one hand, the E.U. seeks to impose a short lifespan for query data.¹⁰ On the other hand, their overview directives related to all data protection and information privacy make exceptions for national security, criminal investigations, or data that are important for scientific, statistical, or historical reasons.¹¹ The E.U. policy on anonymization ignores the value that personalized search query data contain in these areas, and stands in contradiction to the values recognized in the overview language.

4 Several references are made below to WP148, a document, from the so-called Article 29 Working Party in the E.U., which focuses on search query anonymization:
http://ec.europa.eu/justice_home/fsj/privacy/docs/wpdocs/2008/wp148_en.pdf.

5 <http://www.youtube.com/watch?v=JNu1OtkWrOY>. For more details see “Using data to help prevent fraud”:
<http://googleblog.blogspot.com/2008/03/using-data-to-help-prevent-fraud.html>; “Using log data to help keep you safe”: <http://googleblog.blogspot.com/2008/03/using-log-data-to-help-keep-you-safe.html>.

6 <http://www.google.org/flutrends/>

7 “Detecting influenza epidemics using search engine query data,” Ginsberg et al., *Nature* 457, 1012-1014 (19 February 2009); <http://www.nature.com/nature/journal/v457/n7232/full/nature07634.html>.

8 For example, San Francisco, last 7 days:
<http://www.google.com/insights/search/#geo=US-CA-807&date=today%207-d&cmpt=geo>.

9 See, e.g., http://en.wikipedia.org/wiki/Guns,_Germs,_and_Steel and <http://www.pbs.org/gunsgermsteel/about/index.html>.

10 WP148, Section 8: “the Working Party does not see a basis for a retention period beyond 6 months.”

11 Directive 95/46/EC, Section VI, Article 13(1).

Anonymization Is Unnecessary

Under an assumption that the data are useful in some way, are there frameworks for protecting intact search query data so that anonymization is unnecessary? Like health data, I would assume 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, I would differentiate between anonymous storage and anonymous access. Like electronic communications, I would assume that a warrant should be required for unauthorized access. There could be criminal liability for knowingly giving data to unauthorized persons. I also propose that “identifying” information be physically and logically segregated from the rest of the data. The point is 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 a historical record that is invaluable. By defining a reasonable data protection standard on search query data, search engines would be allowed to keep the data if they view it as sufficiently valuable to warrant the protection costs. That is likely a proper boundary between governmental regulation and market forces.

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

Another reason that deletion 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 recognize or acknowledge. They become reliably identifying only in cases of a registration, something that users can choose to avoid.

IP addresses alone are frequently not sufficient to reliably identify a user for at least two reasons: multiple users can share a single IP address (e.g. a public computer in a library), and a single user may use multiple IP addresses (e.g. someone using the same laptop from different locations). 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:
#1 <http://www.geobytes.com/IpLocator.htm?GetLocation>
#2 <http://www.melissadata.com/lookups/iplocation.asp>

For a Stanford University IP Address [171.64.1.134]:
#1 came back with Palo Alto, CA with a 99% certainty.
#2 came back with Stanford University.

For a U.C. Hastings IP Address [209.233.180.24]:
#1 came back with Oakland, CA with an 88% certainty.
#2 came back with San Francisco.

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). The E.U. has made a proposal that IP addresses be

used to allow users to access and/or delete their prior queries.¹² However, assuming that an IP address safely identifies an individual would expose more personal information than would be protected by deleting old IP addresses.¹³ IP addresses and user cookies are simply not, in and of themselves, reliably personally identifying. 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, it should 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 username and password. 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. Anonymization is unnecessary.

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, and with such solutions, it is largely unhelpful.

Many of the issues related to user privacy relate to the web in general and it does not make sense to try to fix them website by website or service by service. These issues include the following:

- cookie use, management, and transparency (best handled in the browser)
- potential abuse of subpoenas, both 3rd party and governmental¹⁴
- 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.¹⁵ 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, it may better serve to argue simply that (any) retained data need to be

12 WP148, p. 23

13 Imagine the owner of an IP address accessing searches of users behind a firewall, such as a cafe, a business, or a school.

14 See, e.g., *Warshak v. U.S.*, 532 F.3d 521, 527 (6th Cir. 2008) (en banc).

15 See, e.g., <http://www.nytimes.com/2006/08/09/technology/09aol.html>.

given proper protection, rather than assuming that anonymization solves the problem, or helps more than incrementally. Since protection must be in place anyway (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 engine’s privacy policy, which may not be immediately apparent.¹⁶

These problems are endemic on 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 get 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 non-registered cookies stored in search query logs.¹⁷ 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 overuse or abuse of subpoenas. However, 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 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 at-

tempts at anonymization. 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.

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 to the market. 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.”¹⁸

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 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 im-

¹⁶ WP148, Section 5.3 (emphasis added).

¹⁷ See, e.g., WP148 Section 3, regarding “flash cookies”.

¹⁸ <http://abcnews.go.com/WNT/story?id=1526798>.

posing a particular solution, anonymization, on all the participants.

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. In particular, certain arguments in the E.U. lack logical consistency and there is a risk that they may influence policy decisions in the U.S. The E.U. requires that data be collected for a stated purpose, and not be used for other or expanded purposes.¹⁹ The problem with this 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 reposing.

Whose Data Are They Anyway?

If governments impose data deletion as a privacy solution, should they have to compensate the search engines for the destruction of a valuable asset? 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? 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 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?²⁰ 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 with 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. In *Ruckelshaus v. Monsanto Co.*, 467 U.S. 986 (1984), the U.S. Supreme Court held that the uncompensated taking by the EPA of a trade secret was unconstitutional. Allowing for data protection, rather than forced deletion, avoids this problem altogether.

Queries as User-Generated Content

Ironically, even the assumption that the user (co-)owns the data does not remove the problems caused by forced 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 (greatly abridged):

```
tuskegee airmen
tuskegee airmen
<http://www.tuskegeeairmen.org/>
<http://en.wikipedia.org/wiki/Tuskegee_Airmen>
tuskegee airmen religion
tuskegee airmen jewish
<http://www.mlive.com/news/kalamazoo/index.ssf/2009/01/unexpected_opportunity_tuskege.html>
<http://ems.gmnews.com/news/2009/0225/bulletin_board/010.html>
tuskegee airmen biography jewish
<http://www.encyclopedia.com/doc/1E1-DavisBOJr.html>
<http://blackhistorypages.net/pages/tuskair.php>
```

¹⁹ WP148, Section 5.2.

²⁰ See, e.g., *Privacy As Intellectual Property?* by Pamela Samuelson http://people.ischool.berkeley.edu/~pam/papers/privasip_draft.pdf

<<http://newpittsburghcourieronline.com/articlelive/articles/39948/1/Tuskegee-Airmen-facts/Page1.html>>
tuskegee airmen religion/religious statistics
<<http://www.sandomenico.org/page.cfm?p=921>>
<<http://tuskegeearmen.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,M1p.68 (last visited Mar. 7, 2009).

[moved to Google book search]
tuskegee airmenlairman jewljewishljudaism

“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_AIQTS1Yy3Dg&id=LY9TAAAAMAAJ&dq=tuskegee+airmenlairman+jewljewishljudaism&q=tuskegee+airmenlairman+jewljewishljudaism&pgis=1#search_anchor>

Suppose that a search engine wanted to publish a book of their 1000 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 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 (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.

Moral rights are inherent in 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.²¹ International agreements under TRIPS (optionally) extend these rights to other copyrightable works.

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 it.

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.²² They become a series of independent elements -- like cutting up poetry into individual phrases. Even simple analyses such as discovering common term replacements or 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.

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,

²¹ 17 U.S.C. § 106A.

²² 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.

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.