

## Internet and Privacy: Toward the Complete Anonymization of Personal Data?



By **Jonathan AMSELLEM**

Project Manager at the Thomas More Institute

The issue of who controls citizens' personal data, however complex and obscure it may often seem, is one of the major challenges facing modern democracy. The heated controversy recently sparked in France surrounding the contemplated Edwige database of records suffices to illustrate this. But the most sensitive issue today involves the stakes represented by the Internet. The development of new information technologies has considerably increased our exposure on this global virtual forum.

The issue of the collection, use and storage of personal data on the Internet has been one of the most sensitive points since the opening up of the WWW to the general public in 1995... 1995-2008: it will effectively have taken thirteen years before the problems in connection with the storage of personal data on the Internet came to be really closely examined. Nothing has really changed since the mid-1990s – i.e., even before the creation of Google, the leading search engine globally! For now, users only have vague and often obscure information available to them concerning the data collected and their use by Internet majors.

### The G29 Recommendation Is Welcome but More Is Needed

In Europe there is a body – with an exclusively advisory role – in charge of advising the European Commission "on any Community measures affecting the rights and freedoms of natural persons with regard to the processing of personal data and privacy": the "Article 29 Data Protection Working Party", or the G29 (1). The G29 brings together representatives from all of the data protection bodies appointed by the EU member states and is currently chaired by Alex Türk, the chairman of the CNIL in France.

An April 2008 opinion of the G29, which went relatively unnoticed, makes real strides in terms of personal data protection: it unanimously recommended limiting to a maximum of six months, compared to the current retention period of twelve months, the period during which search engines may use or store personal data. While this notice is not binding, personal data protection depending solely on national laws, it is an encouraging first step. This is the first time that a recommendation by the G29, which previously issued opinions focusing more on data content, is now taking on the issue of storage periods.

Progress still remains to be made in terms of European coordination. Under the current rules, it is the place of data storage that determines what national law applies. These archaic rules may amuse specialists but no longer have any meaning since data is no longer stored on paper or on disks, but on servers located in different countries...

To bring them up to date, the G29 first identified what type of personal data search engines can obtain and what they can do with that data.

### Personal Data: Toward More and Faster Anonymization

In defining what personal data is, it is appropriate to consider all of the means of identification available to the person responsible for the data processing (data controller), so as to determine whether a person is identifiable or not. According to the CNIL (France), "personal data consists in all anonymous data whose cross-correlation enables identifying a specific person".

The first information a search engine obtains is the IP (Internet Protocol) address. This code is not dissociable from the Internet connection. It is only by this address that websites identify every Internet user – or to be more specific, every computer. This "anonymous" address corresponds to a series of digits (four digits comprised between 0 and 255). But access providers know who is behind an IP address, and it is possible for a third party to correlate the information and determine to whom a profile corresponds. Let's take the example of an ordinary email box: when the user signs up for it, he or she provides specific information concerning his or her identity, gender, home address, etc. This information is directly linked to the IP address and the host knows that the connected IP address corresponds to such and such email address, the gender of the person, their location...

This is how the search engine identifies a user and continually records the clicked links, the sites visited, ads viewed, as well as the key word searches made, frequency of use or visits to a given site. It correlates all of this information to the user IP's address thanks to cookies placed on that person's computer whenever he or she logs on to the Internet. Cookies are connection logs; they enable the websites to identify the users and record their "itinerary".

For the leading search engines such as Google, Yahoo and Microsoft, the collection of such data is the best way of learning about the user's expectations, improving the performances and services offered by their tools, but sometimes also to respond to certain security requirements. These engines basically live off of advertising, and the data harvested constitutes a real goldmine for the targeting of Internet users and serving them with ads tailored to them. Such data also serve to improve the relevance of the contents found. Depending on the frequentation (number of clicks) between one or another site, the number of links and keywords contained by them, search results for a same keyword can change from one day to the next. For such Internet majors, improving the use of their tools thus necessarily involves the collection of personal data. But there are also other reasons for obtaining such data. Legal obligations exist, as do security problems or cases of fraud which require the use of such data. The authorities may use such data in the context of investigations in racism, terrorism, pedophilia cases, etc.

This is why it is difficult to stem the enrichment of personal data and why the G29 decided to only seek data anonymization after six months and not their deletion.

### Microsoft: choice of full anonymization

So how have the three main search engines worldwide reacted to this recommendation by the G29? Microsoft was the first to make known its reaction - which was positive -, on December 9, 2008, indicating that it was willing to follow the G29's recommendation and cut the length of time when data is used from 18 to 6 months before anonymizing them – while specifying that this would

only be useful and effective "provided others followed suit" (2). Microsoft's *Live Search* engine only accounts for 2% of online searches in Europe. The following December 19<sup>th</sup>, Yahoo announced that it would cut its data storage period to 90 days (3 months). Only Google, despite being the leader, has yet to react.

Microsoft did, however, rightly observe that the method enabling the deletion of personal data is in fact more important than the length of the storage period itself. The firm founded by Bill Gates has asked the main search engines to come to an agreement on an ambitious policy on deletion. In effect, Yahoo has declared that full data anonymization after 90 days is not possible for various reasons, both technical and legal. The firm has gone on to declare that it would only delete a part of the address.

This point, apparently complex, raises a new and simple question: does the partial deletion of data prevent data being correlated? The answer is no. By comparing an IP address to a phone number, half of the digits can serve for identification purposes. To take a concrete example: half of the digits of a French phone number, let's say 01 45 55..., suffices to know where the person lives: 01 for Ile-de-France, 45 for the left bank of Paris. Just that small amount of information is enough to learn about the person. This is why Microsoft's declared objective of full data anonymization with the full deletion of IP addresses, despite a storage period cut less than Yahoo's, seems qualitatively more satisfying.

By so doing, Microsoft now appears as the search engine most cooperating with the G29 and as being a step ahead of the "next reform": that of "affirmative consent" by the user. What is involved is to give the user more control by obtaining his or her consent before collecting data that might be "sensitive", meaning data enabling that person's identification. The battle for personal data control on the Internet is a long haul...

### Jonathan AMSELLEM

- 
- (1) Its organization and tasks are laid down in Articles 29 and 30 of Directive 95/46/EC, after which it is named, and by Article 14 of Directive 97/66/EC.
  - (2) John Vassallo, Microsoft advisor for European affairs, in "Microsoft and Google tussle over search anonymity", *The Guardian*, December 8, 2008.
-