International Working Group
on Data Protection
in Telecommunications

675.46.32

**Working Paper and Recommendations
on the Publication of Personal Data on the Web,
Website Contents Indexing and the Protection of Privacy**

53rd Meeting – Prague, 15-16 April 2013

1. Background

One of the main pillars of data protection has always been the data subjects' right to control their own data. An essential element of this control is the right to have one's data deleted if they are processed illegally or if the data subject no longer consents to their processing. The recent proposal by the European Commission for a new regulatory framework tries to strengthen this right by providing for a "right to be forgotten" by others, and on the Web. This is without prejudice to those cases where there is a legitimate and legally justified interest to keep data published and visible, such as in media archives or for the purposes of historical records, and it is clear that the right to be forgotten cannot take precedence *a priori* over freedom of expression or freedom of the media[1].

In view of the present structure of the Web, many issues with respect to how such a "right to be forgotten" could be implemented are still unsolved on the technical as well as on the legal side. Personal data (and any other information), once published online, will very likely remain publicly available. Even if they are deleted on the original website, they may have been linked to or mirrored on other sites before deletion. The Web does not know "how to forget" and there is no simple technical tool available at present which could ensure the systematic deletion of data on the Web (i.e., which could teach the Web how to forget). In short, there is no "erase button" and it is doubtful whether there will ever be one.

However, there are ways to protect the individual's right to be forgotten to a certain extent even today by leveraging tools available to website administrators[2] to limit the exposure of personal information on the Web as well as by making use of/harnessing the power of search engines. On the

[1] The EU Data Protection Reform 2012: Making Europe the Standard Setter for Modern Data Protection Rules in the Digital Age, Viviane Reding SPEECH/12/26, Innovation Conference Digital, Life, Design Munich, 22 January 2012, http://europa.eu/rapid/pressReleasesAction.do?reference=SPEECH/12/26; for a critique of this approach see Rosen, The Right to Be Forgotten, 64 Stan. L. Rev. Online 88

[2] One such set of tools are the Google Webmaster Tools, which allow webmasters to see how Google crawls and indexes their site and allows webmasters to influence how the URLs that are indexed are displayed. A link to the tools is available at http://www.google.ca/webmasters/.

Secretariat
Berliner Beauftragter für
Datenschutz und Informationsfreiheit
An der Urania 4- 10
D-10787 Berlin
Phone +49 / 30 / 13889 0
Fax: +49 / 30 / 215 5050

E-Mail:
IWGDPT@datenschutz-berlin.de

Internet:
http://www.berlin-privacy-group.org

The Working Group has been initiated
by Data Protection Commissioners
from different countries in order
to improve privacy and data protection
in telecommunications and media

current Web, the right to be forgotten[3] might better be interpreted and implemented as a "right not to be found".

2. The prospects of regaining control of the exposure of personal data on the Web

The increasing publication of personal data on the Web over the past year has given rise to new challenges and risks for the privacy of citizens, while aggravating existing risks at the same time. The advent of social networks has played a particularly crucial role in this context[4].

While technologies fostering the publication and making available of data – including personal data – on the Web have made dramatic progress in this context, the development of technologies to control the availability of such data on the Web still seems to be in its infancy. While work on a "policy-aware web"[5] has taken place over the past decade, we still seem to be far from any effective, easy-to-use and widely available tools which would enable citizens to (re-) gain even a limited amount of control over their data once they have been published on the Web.

One possible design goal for such technologies could be fostering the deletion of all copies of that data from any device or storage area where it is retained. At present, this may well pose problems of scalability (even with an automated approach), especially if specific data has been disseminated on the Web or further elaborated or re-contextualized over time by the community of Web users. There is currently no technical way to identify and locate all copies of an item and copies of information correlated with that item on the Web. However, this may be possible in a future "policy-aware Web".

For newly generated data, exposure on the web could be limited by setting time limits (expiration dates) on the given item. This can be accomplished in many ways. For instance, one might equip data with "active" (executable) software which intervenes when the expiration deadline is reached in order to disable data display on a screen, or disable the ability to make screenshots of the image, or ultimately to delete or encrypt the original content. Alternatively, data can be "tagged" with an expiration date, so that all servers handling that item can take account of that date and remove the data after the expiry date.

Further interesting examples of how to customize the lifetime of newly generated data on the Web are given by some other emerging applications. For instance, users may utilise a secure overlay network restricting visibility of content, such as a post or image, to a community belonging to the same overlay network by using end-to-end security and access control policies. In yet other applications, a mobile text message remains available to a user until a configurable expiration date. Finally, reference can be made to "user centric" solutions, where the legitimate owner of a data may selectively provide access to it, releasing links to the place where the data is actually stored only within a specified timeframe.

These examples may serve as building blocks for a future "policy-aware Web". However, a lot of thorough research and development is necessary to further develop these elements to be effective tools for better protecting the privacy of citizens. The Working Group calls upon the relevant actors in

---

[3] Note that the term „right to be forgotten" is used in a broader sense in this paper than in the proposed EU Privacy Regulation, and that this paper does not make any statement on whether a "right to be forgotten" can be implemented in that regulation or not.

[4] Cf. the Report and Guidance on Privacy in Social Network Services – "Rome Memorandum" of this Group – (Rome (Italy), 3./4.03.2008); http://www.datenschutz-berlin.de/attachments/897/675.36.5.pdf

[5] For some existing proposals for creating a "policy-aware Web" cf. footnote 27 on page 10 of the „Rome Memorandum" (footnote 4 above). The concept of the policy-aware Web combines several existing technologies, namely structured data, identity management, access control, and sticky policies (i.e., use policies that travel with the data itself).

this field (Industry, academia, and governments) to further strengthen their efforts to make progress in this field.

3. Restricting availability of personal data on the Web by controlling their indexability by search engines

Another building block for restricting availability and contributing to erasability of data on the current Web is to restrict its availability in the results of queries to search engines[6]. This is already technically feasible and available as an option to website administrators. It essentially relies on two alternatives: the robots.txt protocol[7], and the use of "tags" attached to an item to signal that a specific content or web page should not be indexed by a search engine.

The robots.txt protocol works by way of a small set of simple instructions coded in a text file (the robots.txt file) placed in the root directory of a domain (e.g., http://example.com/robots.txt). This file is read, if present, by a crawler (software program used by search engines to give a "snapshot" of a website) prior to indexing the relevant website. The instructions in question allow requesting *specific crawlers* to ignore *specific files and/or directories* in the website. The instructions are executed by crawlers after text matching of alphanumerical strings according to the sequence followed by the instructions in the robots.txt file. Limitations of the protocol include a lack of sufficient scalability, it does not work with ftp-servers and the information is lost when content is copied from a website[8].

Alternatively, different categories of "tags" can be used as attributes of a specific web page (but also in connection with individual elements of a specific page, such as an image or a file therein) to signal that the item/page should not be included in the results of a search query.

It should be emphasized that these approaches are both entirely based on net etiquette (i.e., on the co-operation of the parties involved). As such, they are very difficult to enforce. Their implementation by websites, and adherence to by search engines, is strictly voluntary. Thus, while they can mitigate the risk of indexing determined by linkage from third party sites, they cannot ensure per se that a given item of information will never be indexed by a search engine, in particular if that item is publicly accessible and can be processed by other websites with different crawler access rules.[9]

4. Recommendations to Website Administrators

Website administrators play a crucial role in both categories of erasability described above, namely through their capabilities for limiting the exposure of data and restricting the indexability of items. In order to contribute to the goals set out above, the Working Group makes the following recommendations:

- Website operators should inform their users about what personal data they retain and for what purposes. They should provide their users with an easy mechanism to access their per-

---

[6] See also Recommendation CM/Rec(2012)3 of the Council of Europe on the protection of human rights with regard to search engines.

[7] The robots.txt protocol is also referred to as the Robots Exclusion Protocol and the Robots Exclusion Standard. The protocol is defined in an expired IETF Internet Draft, available online at http://www.robotstxt.org/norobots-rfc.txt.

[8] Changes in web content and/or indexing preferences can sometimes not be reflected in search results as well. Getting search engines to update their indexes when sites change has proven to be a significant problem.

[9] In this regard, see also the Recommendations contained in the IWGDPT's "Common Position on Privacy Protection and Search Engines" as adopted in 1998 and revised in 2006; http://www.datenschutz-berlin.de/attachments/238/search_engines_en.pdf .

sonal data, and allow them to correct and/or to delete them permanently, as provided for by existing privacy legislation. Such access mechanisms should be user friendly and should not result in any additional cost to users or impose unjustified delays or operational burdens.

- Upon a data subject's specific request, and if no other legitimate interests or legally binding constraints exist, webmasters should promptly remove the relevant piece of information from their website. In addition, they should signal to search engine providers to re-index the respective part of the website, in order to have the data also deleted from the search index and any cache copies of the search engines.

- Webmasters should provide their users with specific tools to allow customizing their search indexing preferences[10]. Alternatively, consideration could be given to using the "noindex" meta-tag – to be included in the HTML code of the relevant page or in the HTTP header – or the sitemap.xml file to signal the relevant search preferences in connection with specific items[11].

- Special care should be taken in writing the robots.txt file as regards the lexical and semantic correctness of the instructions as well as their inherent logical consistency (to prevent conflicting and/or overlapping instructions). It should be pointed out that *failing specific exclusion instructions* in the robots.txt file, a crawler will assume that the administrator allows site indexing or the indexing of specific sub-directories (i.e., a crawler will assume that website contents should be made available to search engines).

- It should be observed that the robots.txt protocol does not lend itself to regulating access to especially "risky" contents such as traffic data generated by electronic communications services, SMS-message contents, voice mail storage, location data, financial data etc., nor is it aimed at preventing access to specific administrative areas in a website.[12] The robots.txt protocol does not replace cryptography or access control mechanisms.

- If a webmaster intends to signal that specific pages and/or files should not be indexed by search engines, special care should be dedicated to the selection of URLs. Indeed, since the robots.txt file is publicly visible, relying on "self-explanatory" URLs might ultimately enhance exposure of the relevant contents and thereby void the benefits of the protocol. The contents

[10] Cf. the mechanism provided by the "blogger.com" platform enabling users to set up their indexing preferences in a specific form to be filled when subscribing to the blog service, instructing the webmaster on how to configure his own robots.txt file (http://buzz.blogger.com/2012/03/customize-your-search-preferences.html).

[11] This recommendation is especially relevant in dynamic environments or complex websites, where the robots.txt solution might not scale enough with the size of the website. An example of the use of the robots.txt commands to signal a search engine the expiration date of a page may be found at http://googleblog.blogspot.fr/2007/07/robots-exclusion-protocol-now-with-even.html. Similarly, the sitemap.XML file signals how frequently a web page may change, and the priority level that the webmaster attributes to a URL, allowing the search engine to potentially select the appropriate refresh rate. Cf. also http://lists.w3.org/Archives/Public/public-privacy/2012OctDec/0224.html .

[12] In July 2011, about 8000 SMS messages received on the mobile network of MegaFon were indexed by Yandex, a Russian search engine, making content data and addressees' mobile phone numbers publicly available. Similarly, 43,000 SSNs belonging to Yale University students were disclosed in August 2011 after being indexed by Google because they had been stored in a public subdirectory of an ftp server.

of the robots.txt file are especially valuable to hackers as well as to any entity seeking to disseminate/acquire personal data.

5. Recommendations to Search Engines

As one of their core activities, search engine providers work mainly as information brokers/intermediaries[13]. However, there are also certain types of processing in respect of which they act as separate data controllers.

In particular, some search engines perform many different activities, ranging from indexing of websites, to storing the respective contents temporarily to enable users' retrieval of information in case a server and/or link is down/unavailable. This caching constitutes a re-publication for which the provider of the search engine is deemed to be a data controller[14].

Accordingly, the following recommendations for search engine providers are distinguished according to the different roles played by them.

Mere Indexing

- Search engines should always respect indexing preferences expressed by websites with respect to the content they host, whether via the robots.txt file or via other "noindex" tagging mechanisms, including expiration date commands. Such indexing preferences can be expressed before the first crawling of the website or once it has already been crawled. In the latter case, updates on the indexing performed by a search engine should be carried out as soon as possible.

- Search engines should enhance the effectiveness of their communication channels with webmasters in order to be notified rapidly of any change in the indexing preferences, expressed by webmasters by means of the appropriate commands of the robots.txt protocol, or any modification of items within a website. The update/rectification procedures should be as privacy-friendly as possible – in particular, no additional personal data should be required from users that request certain items of personal information to be updated/rectified.

- Search engines should adjust their crawling rate according to the search preferences expressed by webmasters. They should also execute any requests by webmasters for re-indexing their websites or parts thereof following the deletion or correction of personal data without undue delay.

---

[13] Cf. the Opinion on data protection issues related to search engines (WP148) by the Article 29 Working Party of European Privacy Commissioners (http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2008/wp148_en.pdf). Note that this issue is currently before the European Court of Justice.

[14] As pointed out in the Opinion on search engines by the Article 29 Working Party of European Privacy Commissioners (WP148), "… any caching period of personal data contained in indexed websites beyond (…) technical availability, should be considered an independent republication. The [Article 29] Working Party holds the provider of such caching functionalities responsible for compliance with data protection laws, in their role as controllers of the personal data contained in the cached publications."

- Since there is as yet no consistent interpretation by search engines of the instructions written in a robots.txt file or in other indexing preferences signaling mechanisms (e.g., metatags, sitemap.xml files), it is difficult to foresee what impact such mechanisms will have on indexing of a website by the different crawlers. It is desirable that search engines agree on a "modus operandi" in this regard. The mechanisms applying to the individual instructions should be described clearly on a page that should be easily accessible by users (e.g. from the main pages of the search engine portals).

- Search engines should be involved to a greater extent in supporting website administrators by providing tutorials and/or tools for the automated analysis of indexing preferences. This will enable administrators to check what effects the instructions they are giving will produce in terms of indexing.

- Search engines should more clearly specify timing and criteria of the "crawling" they perform on a given website, so that administrators and users can reasonably gauge how long a given piece of information will remain available as a search result.

Temporary Storage of Crawled Information

- Search engines should implement specific crawlers if they intend to group data according to different categories and for different purposes (e.g., general indexing, news, images, etc.) in order to allow website administrators to better control the context in which information will be published.

- When indexing a website, search engines should accept more complex and more granular instructions for their crawlers such as the following:
    - Permissions to index information for specific purposes (e.g., general-purpose search engines vs. news search engines, etc.)[15];
    - Permissions to temporarily store information for specific purposes, including the respective time limits (e.g. caching, snippets);
    - Permissions to communicate information to third parties for specific purposes;
    - Permissions to process the retrieved data for specific use-cases[16] based on the occurrence of features such as geographic area or IP address ranges.

- Where crawling is followed by temporary storage of site contents for purposes other than that of enabling users to access those contents in the event the given server/network is down/unavailable, search engines should provide site administrators with clear-cut, specific information on the timeline and technical mechanisms applying to said storage.

---

[15] Cf. e.g. the findings reported by the Italian Antitrust Authority following a complaint lodged by the Italian Federation of News Publishers against Google. Thereafter Google committed itself publicly to complying with a set of undertakings so as to provide publishers with tools that should help them distinguish between indexing of contents on the generalist search engine and indexing on the news search engine.

[16] Given the increasingly complex use-cases that apply to the information crawled by search engines, it might be appropriate to consider reversing the current pattern, whereby crawlers are allowed to read information if an instruction is formally incorrect or cannot be interpreted by the crawler. If it proves impossible to interpret a complex set of instructions, the latter should be interpreted by default as a ban on indexing/storage by the crawler.

- Search engines, upon specific requests issued by webmasters through their search preferences, should promptly delete any cached copy of the data retrieved from websites, and should abstain from further processing these data, mitigating in this way the risk of data dissemination and overexposure.

6. Final Caveat

In this paper, the Working Group has explored tools for controlling the availability of (personal) data on the Web which are available today to users, webmasters and search engines, mostly based on limiting contents exposure on a website either through the application of (automatic) deletion mechanisms[17] or via the implementation of search preference signaling protocols. It should be recalled that the latter still rely on simple on/off (binary) rules applying to crawlers, and were designed over 15 years ago. Conversely, search engines have become increasingly sophisticated over the years and the rather simple inclusion/exclusion mechanism underlying the protocol in question is no longer fully capable to cope with the ever increasing scope of data retrieval and storage. It should, for instance, be emphasized that the availability of data (including data which users disclose about themselves), in combination with facial recognition techniques and location data, can ultimately allow the indexing of individuals rather than simply of contents, or of information. An urgent focus on these aspects is therefore necessary.

Another prospective technological breakthrough for better protecting personal data on the Web may be the advent of the "policy-aware, semantic Web", where data could be inextricably linked with attributes (e.g., a "meaning") and access rules. This would allow, on the one hand, for the creation of new relations between data and enhance the concept of an interconnected world whilst providing, on the other hand, more effective mechanisms to identify and locate content, and potentially also copies of information correlated with that item based on attribute matching (rather than by simple text matching as it happens today). This would make it conceivable to remove information from a multiplicity of websites and to de-link search results from websites, thus avoiding any unintended data dissemination[18].

Of course, usability of the Web should not be undermined and a balance must be struck between innovation and individuals' fundamental rights to privacy and data protection. Further consideration should be given to the option of introducing more granular mechanisms that are not based on the simple exclusion/inclusion rule, but rather attempt to enable data subjects to better express their own search preferences and link the information to the appropriate context (for instance, by allowing data subjects to signal whether or not a given piece of information is still current or relevant, or the occurrence of any events that may have impacted on that information). This would afford data subjects wider options than the simple choice between blanket overexposure on the Web or a complete abstention from new technologies.

There are major, growing interests of an economic nature vested in both search engines and website administrators pushing for the widest possible availability of data through the implementation of data and information indexing. This indexing of web sites serves the economic interest of certain market

[17] It is worth highlighting that, due to the public nature of the Web, when a website administrator wants to remove content from the "public sphere", other access control mechanisms should be implemented such as user authentication and/or data encryption.

[18] Google recently announced an update to its search algorithm which will lower the ranking of sites with high numbers of removal notices, to be applied only in cases of copyright infringements (http://insidesearch.blogspot.fr/2012/08/an-update-to-our-search-algorithms.html or http://www.google.com/insidesearch/howsearchworks/).

players, and removing publicly available web contents or signaling that such contents should not be indexed and retrieved via a search engine is bound to impact on market dynamics and business models. Co-operation of the various stakeholders is necessary to appropriately reconcile the interests in question with the need for privacy protection.