Do they know me? Deconstructing identifiability

Leenes, R.E.

Published in:
University of Ottawa Law and Technology Journal

Publication date:
2007

Document Version
Publisher's PDF, also known as Version of record

Do They Know Me? Deconstructing Identifiability

Ronald Leenes*

DATA PROTECTION REGULATION AIMS TO PROTECT INDIVIDUALS against misuse and abuse of their personal data, while at the same time allowing businesses and governments to use personal data for legitimate purposes. Collisions between these aims are prevalent in practices such as profiling and behavioral targeting. Many online service providers claim not to collect personal data. Data protection authorities and privacy scholars contest this claim or raise serious concerns. This paper argues that part of the disagreement in the debate stems from a conflation of distinct notions of identifiability in current definitions and legal provisions regarding personal data. As a result, the regulation is over- and under-inclusive, addresses the wrong issues, and leads to opposition by the industry. In this paper I deconstruct identifiability into four subcategories: L-, R-, C- and S-identifiability. L-identifiability (look-up identifiability) allows individuals to be targeted in the real world on the basis of the identifier, whereas this is not the case in the other three. R-identifiability (recognition) can be further decomposed into C-type (classification) identifiability, which relates to the classification of individuals as being members of some set, and S-type (session) identifiability, which is a technical device. Distinguishing these types helps in unraveling the complexities of the issues involved in profiling, dataveillance, and other contexts. L-, R-, and C-type identification occur in different domains, and their goals, relations, issues, and effects differ. This paper argues that the different types of identifiability should be treated differently and that the regulatory framework should reflect this.

LA RÉGLEMENTATION DE LA PROTECTION DES DONNÉES VISE À PROTÉGER LES PARTICULIERS contre le mésusage et l'abus de leurs renseignements personnels, tout en permettant aux entreprises et aux gouvernements de se servir de ces renseignements à des fins légitimes. Les collisions entre ces objectifs sont courantes dans les pratiques que sont notamment le profilage et le ciblage comportemental. Bon nombre de fournisseurs de services affirment ne pas recueillir de renseignements personnels. Les instances responsables de la protection des données et les spécialistes des questions de respect de la vie privée contestent cette revendication ou, en tout cas, émettent de sérieuses réserves à ce sujet. Dans ce texte, on soutient que le désagrément entourant ce débat découle en partie de la méthode d'appariement de notions distinctes « d'identifiabilité » dans les définitions actuelles et les dispositions législatives relatives aux renseignements personnels. Par conséquent, la réglementation envisagée est à la fois trop et pas assez « inclusive », elle traite les mauvaises questions et suscite l'opposition au sein de l'industrie. Dans ce texte, je déconstruis l'identifiabilité en quatre sous-catégories : L-, R-, C- et S-. L signifie « look-up identifiability » (soit la recherche de l'identifiabilité) et permet aux personnes d'être ciblées dans le monde réel à l'aide d'un identificateur, alors que ce n'est pas le cas des trois autres sous-catégories. En effet, R signifie « identifiability » dans le sens de la reconnaissance et peut à son tour être décomposée en une sous-catégorie de type C (pour classification), laquelle réfère à la classification des individus en tant que membres d'un ensemble et une autre sous-catégorie appelée S (pour session) correspondant à une aide technique. Établir une distinction entre ces divers types permet de mettre en lumière les complexités des questions en jeu dans le cadre du profilage, du contrôle des données et d'autres contextes. L'identification des types L-, R-, et C- se produit dans différents domaines et leurs objectifs, leurs rapports, les questions en jeu et leur incidence diffèrent de l'un à l'autre. Dans ce document, on soutient qu'il faudrait traiter de manière spécifique chacun des différents types d'identifiabilité et que le cadre réglementaire devrait refléter cette réalité.

* Full Professor in Regulation by Technology, TILT—Tilburg Institute for Law, Technology, and Society, Tilburg University, The Netherlands. The author greatly acknowledges Teresa Scassa, Bert-Jaap Koops, Anton Vedder, Bart Custers, Jane Bailey, Tal Zarsky and the external reviewers for their comments on drafts of this article. This paper was written during the author's stay at the University of Ottawa Law and Technology Group and the Canadian Internet Policy and Public Interest Clinic (CIPPIC).The author is indebted to Ian Kerr and Pippa Lawson for providing the environment to work on the paper and Apple's investors for the financial support they indirectly provided.

Do They Know Me? Deconstructing Identifiability

Ronald Leenes

1. INTRODUCTION

THE "REVEALED I" CONFERENCE[1] FEATURED A DEBATE between a representative of the Internet Advertising Bureau and privacy advocates about some of the pressing privacy issues of contemporary internet use: behavioral targeting and profiles.[2] While the topic in itself is very interesting and important, the discussion also clearly showed a conceptual confusion that is present in many current discussions about data protection and online privacy. In the context of behavioral targeting, the confusion amounts to something like this. We (privacy advocates) are concerned about the profiling and behavioral targeting conducted by the advertisement industry on the basis of the online behavior of individual internet users. The advertisement industry counters that although one may find profiling and behavioral targeting troublesome, we (the advertisement industry) do not collect personal data,[3] and hence we consider ourselves to operate within the boundaries of the law (if there is one), so where is the problem?

The problem in this line of argument by the advertisement industry is that it implies a very shallow definition of identifiability. Everyone agrees that collecting names and addresses of internet users clearly amounts to collecting personal data and that this data identifies individuals. Most service providers are aware that the processing of this kind of data requires care, which involves certain obligations in some jurisdictions, such as the European Union (EU). At the other end of the spectrum, there is data that clearly does not pertain to individuals

1. Organized by the "On the Identity Trail" Project, University of Ottawa (26–27 October 2007), <http://idtrail.org/>.
2. Behavioural targeting was not only an item on the agenda of the "Revealed I" conference. A couple of days later, on November 1 and 2, 2007, the United States Federal Trade Commission (FTC) hosted a Town Hall entitled "eHavioral Advertising: Tracking, Targeting, & Technology," which brought "together consumer advocates, industry representatives, technology experts, and academics to address consumer protection issues raised by the practice of tracking consumers' activities online to target advertising, or 'behavioral advertising.'" See Federal Trade Commission, "eHavioral Advertising: Tracking, Targeting, & Technology," <http://www.ftc.gov/bcp/workshops/ehavioral/index.shtml>.
3. During the FTC Town Hall, several of the industry representatives reiterated that they do not collect personally identifiable information (PII). For a transcript of their statements, see FTC Office of Public Affairs, "eHaviorial Advertising: Tracking, Targeting, & Technology" (1-2 November 2007), <http://htc-01.media.globix.net/COMP008760MOD1/ftc_web/FTCindex.html#Nov1_07>.

and the collection of this non-personal data does not impose such obligations and care. An extreme example can be offered: few people would consider that collecting weather data introduces privacy issues. Between these extremes there are kinds of data for which it is less clear whether they constitute personal data; for instance, are Internet Protocol (IP) addresses personal data? The short answer is that this is not entirely clear.[4]

What is certain is that identifiability goes well beyond names and addresses. Most people immediately know who I mean by "the guy with the reindeer who visits North America around the end of the year," without having to spell out his name. Therefore, insisting that data collection is unproblematic if it does not involve personal data is misleading because it neglects the wider scope of identifiability which lies at the heart of data protection and informational privacy. The advertisement industry's statement that they do not collect personal data may be plain rhetoric, but it may also signify that the question as to what amounts to personal data and *identifiability* in the online world is debatable.

In this paper I will argue that the notion of "identifiable person" in current legal provisions and definitions conflates a number of distinct types of identifiability that are best distinguished to prevent the kind of discussions described in the introduction. Deconstructing the concept of "identifiable person" will help in singling out the various kinds of privacy issues associated with web browsing and will facilitate defining measures to more effectively address the issues. One of the results of such an exercise may be that privacy advocates and the "industry" can move closer, even though they may have different interests at the end of the day.

Let us examine the issues surrounding identifiability. I have a European background and, consequently, this article will focus mainly on European terminology and European regulation; however, the point I try to make is general and has equal merit for North American debates.

*

2. PERSONAL DATA

DATA PROTECTION REGULATION ADDRESSES the proper use of personal data.[5] Therefore, a central concept in the European Directive 95/46/EU (generally known

4. In Europe, the Article 29 Data Protection Working Party considers IP addresses to be personal data in most cases: "Internet access providers and managers of local area networks can, using reasonable means, identify Internet users to whom they have attributed IP addresses as they normally systematically 'log' in a file the date, time, duration and dynamic IP address given to the Internet user. The same can be said about Internet Service Providers that keep a logbook on the HTTP server. In these cases there is no doubt about the fact that one can talk about personal data in the sense of Article 2 (a) of the Directive […]." European Commission, Article 29 Data Protection Working Party, "Opinion 4/2007 on the concept of personal data," at p. 16, (20 June 1995), <http://ec.europa.eu/justice_home/fsj/privacy/docs/wpdocs/2007/wp136_en.pdf> [Opinion 4/2007]. A different position is that of the Hong Kong Privacy Commissioner: "An Internet Protocol (IP) address is a specific machine address assigned by the web surfer's Internet Service Provider (ISP) to a user's computer and is therefore unique to a specific computer. An IP address alone can neither reveal the exact location of the computer concerned nor the identity of the computer user. As such, the Privacy Commissioner for Personal Data (PC) considers that an IP address does not appear to be caught within the definition of 'personal data' under the PDPO." Press Releases, "LCQ17: IP addresses as personal data," (3 May 2006), <http://www.info.gov.hk/gia/general/200605/03/P200605030211.htm>, as quoted on Google's Global Privacy Counsel Peter Fleischer's blog, Peter Fleischer, "Privacy…?" (5 February 2007), <http://peterfleischer.blogspot.com/2007/02/are-ip-addresses-personal-data.html>.

5. See for instance Preamble 10 of the European Community, *Council Directive 95/46/EC of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data*, <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:EN:HTML>, [1995] *Official Journal of the European Union* L281, at p. 31 [Data Protection Directive].

as the Data Protection Directive, or DPD) is "personal data," which according to Article 2(a) means "any information relating to an identified or identifiable natural person ('data subject') […]."[6] Contrary to other jurisdictions, such as Canada, which leave the concept of "identifiable person" open to common sense interpretation and case law, the DPD provides some guidance as to what identifiable means in the data protection context through Article 2(a). An "identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity."[7]

There is much to be said about this provision,[8] but I will be brief here. Article 2 distinguishes between *identified persons*, meaning individuals already singled out in an audience, and *identifiable persons*, reflecting the mere possibility to single out certain individuals in an audience. Identification is therefore a successful attempt to identify an identifiable person. Our principal concern for now is "identifiable" person. A more formal way of defining identifiability is: "Identifiability is the possibility of being individualized within a set of subjects, the identifiability set."[9] The prime characteristic of identifiability is therefore the fact that a person can be individualized (or singled out) in a set of individuals.

There are different ways in which this singling out can be done. One form is having the individual's name (and possibly some additional data), which makes it possible to call out for this individual or look him or her up in some register. My name, Ronald Erik Leenes, should be sufficient to identify me in most audiences because I am fairly certain that I am the only one with this name (in the world). This is certainly the case in smaller identifiability sets. For instance, calling out my name in a University of Ottawa Law & Technology Group meeting will be sufficient to draw my attention and thereby single me out in the group. Having my name should also be sufficient to find my room at Tilburg University by consulting the university's online directory. These two contextual cues should also be sufficient to find out attributes to locate me in other environments, such as crowds (the university's website contains pictures of me) or address me privately (my home address can be found in the phone book).

There are also other forms of identifiability. If the identifiability set is overseeable (for instance, a group of people on a square or in a room), then pointing at a specific individual equally counts as individualizing this person in the set. Most people are perfectly capable of pointing out Santa in an ordinary crowd.[10]

There is also a third option, in which the identifiability set need not be present or known to the observer. In this case, the entity doing the identification[11] has information about attributes of the identifiable person that allows the recognition of this individual should he or she ever appear. A popular use of this kind of identifier is agreeing to wear a distinctive feature (for example, a red scarf or a blue coat) to facilitate blind daters to recognize each other when they first meet in person.

6. Data Protection Directive, *supra* note 5 at art. 2(a).
7. Data Protection Directive, *supra* note 5 at art. 2(a).
8. For example, see Opinion 4/2007, *supra* note 4.
9. Andreas Pfitzmann and Marit Hansen, eds., "Prime: Dictionnary," <https://prime.inf.tu-dresden.de/prime/space/Identifiability>. See also Tu Dresden, Faculty of Computer Science, Institute of Architecture, "Privacy and Data Security," <http://dud.inf.tu-dresden.de/Anon_Terminology.shtml>.
10. Individualizing Santa at the annual Santa convention, also known as the Amalgamated Order of Real-Bearded Santas (AORBS) convention, is a completely different story.
11. Who would confusingly be called the identifier, so let's not use this term in this context.

What these cases have in common is that they use identifiers. The Data Protection Directive acknowledges this and states that identification can be done by means of identifiers, such as identification numbers, or by one or more factors specific to his or her physical, mental, economic, cultural or social identity. This is further explained by the commentary to the Directive which states:

> a person may be identified directly by name or indirectly by a telephone number, a car registration number, a social security number, a passport number or by a combination of significant criteria which allows him to be recognized by narrowing down the group to which he belongs (age, occupation, place of residence, etc.).[12]

The Data Protection Directive therefore distinguishes between *direct* identification (names) as well as *indirect* identification, which relates to the other forms of pointing out individuals, including identifying Santa and the blind daters by referring to their physical appearances.

Indirect identifiers introduce complexity, disputes, and, in any case, questions. For instance, what counts as an identifier? Are identifiers universal or relative to a specific context and its users? What may be a useable identifier in the hands of one person may be useless in the hands of another. For instance, when I make my friend's driver's license number publicly available, then some people (the police for instance) would be able to identify her through this information, but certainly not everyone would be able to.

Identifiers also come in all sorts of shapes with different characteristics. There is a fundamental difference between identifying my friend by her appearance and identifying her on the basis of her driver's license number, which is crucial for the understanding of identifiability. The first kind of information (appearance) allows for recognizing my friend on the street, which is not possible with the second kind of information (driver's license number), unless one is able to inspect people's driver's licenses on the street. The availability of the driver's license number allows for something that is impossible on the basis of appearance data: finding out one's civil identity.

A person's name, or civil identity, plays an important role in the identifiability debate and is, in my opinion, one of the reasons why the debate is so blurry.[13] Let us start with a sensible account of the role of names in identification. In their opinion on "personal data," the Article 29 Working Party on Data Protection writes:

> Concerning "directly" identified or identifiable persons, the name of the person is indeed the most common identifier, and, in practice, the notion of "identified person" implies most often a reference to the person's name. In order to ascertain this identity, the name of the person sometimes has to be combined with other pieces of information (date of birth, names of the parents, address or a photograph of the face) to prevent confusion between that person and possible namesakes. […] The name may also be the starting point leading to information about where the person lives or can be found, may also give

12. Opinion 4/2007, *supra* note 4 at pp. 12–13.
13. Identifiers are closely associated to names. Take for instance the Wikipedia definition of "Identifiers," <http://en.wikipedia.org/wiki/Identifier>: "Identifiers (IDs) are lexical tokens that *name* entities. The concept is analogous to that of a 'name'. Identifiers are used extensively in virtually all information processing […]" (emphasis added).

> information about the persons in his family (through the family name) and a number of different legal and social relations associated with that name (education records, medical records, bank accounts). It may even be possible to know the appearance of the person if his picture is associated with that name. All these new pieces of information linked to the name may allow someone to zoom in on the flesh and bone individual, and therefore through the identifiers the original information is associated with a natural person who can be *distinguished* from other individuals. [14]

Identification that involves the name of the identified is certainly something that has to be taken seriously because it allows tracking down and haunting the identified individual. Therefore, there are sound reasons to regulate this kind of identification as is done in the Data Protection Directive.

Fortunately, the Article 29 Working Party acknowledges that there is more to identification than being able to establish the identified individual's name:

> [W]hile identification through the name is the most common occurrence in practice, a name may itself *not be necessary* in all cases to identify an individual. This may happen when other "identifiers" are used to *single someone out*. Indeed, computerised files registering personal data usually assign a unique identifier to the persons registered, in order to avoid confusion between two persons in the file. Also on the Web, web traffic surveillance tools make it easy to identify the behaviour of a machine and, behind the machine, that of its user. Thus, the individual's personality is pieced together in order to attribute certain decisions to him or her. Without even enquiring about the name and address of the individual it is possible to categorise this person on the basis of socio-economic, psychological, philosophical or other criteria and attribute certain decisions to him or her since the individual's contact point (a computer) no longer necessarily requires the disclosure of his or her identity in the narrow sense. In other words, *the possibility of identifying an individual no longer necessarily means the ability to find out his or her name*. The definition of personal data reflects this fact.[15]

Now although the Data Protection Directive and the Article 29 Working Party do seem to get it right, the idea that identification and having an individual's civil identity (i.e. name) are two separate notions is certainly not common in the real world. Identification is usually associated with obtaining an individual's name, and most cases pertain to this issue.[16] While being able to relate to the identified individual's name and the consequences this may have for this individual—both online and offline—is a genuine concern, I want to argue that we should pay more attention to identification in the broader sense. Many current online privacy concerns relate to situations where the name of the user is not relevant at all. This lack of interest in obtaining the names of the individuals being profiled and targeted by the "industry" may explain why their behaviour so

14. Opinion 4/2007, *supra* note 4 at p. 13.
15. Opinion 4/2007, *supra* note 4 at p. 14 (emphasis added).
16. Examples are the numerous cases where copyrights holders seek to obtain the names of copyright infringers from ISPs on the basis of their IP addresses, such as *BMG Canada Inc. v Doe*, 2005 FCA 193, <http://reports.fja.gc.ca/en/2005/2005fca193/2005fca193.html> and *Irwin Toy Ltd. v Doe* (CAN Ont Sup Ct J, 2000) [2000] O.J. No. 3318.

far has not attracted much attention from legislatures and privacy watchdogs.[17] A data protection focus on preventing names from being collected and used in the online world misses the point. Current online "privacy" issues are much subtler.

*

3. THE IDENTIFICATION INDUSTRY

TO UNDERSTAND WHY IDENTIFIERS SHOULD CONCERN US, let us have a look at one branch of the industry that has an interest in identifying online users: search engines and advertisement serving companies. Search engines are provided by corporations with commercial interests. Their business models are based on providing advertisements to their users. The better these advertisements are tailored to the search engine's users, the more likely the viewers are to follow up on the advertisement[18] and the less annoying these advertisements will be judged by the users.[19] Search engine providers therefore have a clear commercial interest in knowing who their users are. Google's CEO makes no secret of this: "We are moving to a Google that knows more about you."[20] Apart from registered services, such as myGoogle and gMail that require users to provide personal data that connects their online identity to their civil identity, Google also uses indirect identifiers.[21] Google keeps track of the queries submitted by their users and the corresponding search results. A search engine can employ two ways of knowing their users' preferences and habits without requiring them to log in using a username and password. These methods rely on cookies and IP addresses as identifiers.[22]

When a user first contacts the search engine, a cookie will be stored in the user's web browser. A cookie is a small amount of information containing the address of the cookie provider and some additional data in the form of the name of an attribute and its value. Often a cookie will be set containing a unique identifier, but additional cookies may be set containing data such as the last time the site was

17. There have been enquiries by data protection authorities and other oversight committees about cookies. For instance, see European Commission, Article 29 Data Protection Working Party, "Privacy on the Internet—An Integrated EU Approach to On-line Data Protection," (21 November 2000), <http://ec.europa.eu/justice_home/fsj/privacy/docs/wpdocs/2000/wp37en.pdf>. Also, the United States Federal Trade Commission delivered a report on profiling as early as 2000, Chairman Robert Pitofsky et al., "Online Profiling: A Report to Congress," (June 2000), <http://www.ftc.gov/os/2000/06/onlineprofilingreportjune2000.pdf>, as well as organized the November 1 and 2, 2007 Town Hall entitled "eHavioral Advertising: Tracking, Targeting, & Technology," *supra* note 2.
18. Christopher Soghoian, "The Problem of Anonymous Vanity Searches," (2007) 3:2 *I/S: A Journal of Law and Policy for the Information Society,* <http://ssrn.com/abstract=953673>.
19. According to a study carried out by the industry, people prefer relevant advertisements over non-relevant advertisements. See for instance, Mike Walrath, "FTC Town Hall: Behavioral Targeting Today: Understanding the Business and Technology," (1 November 2007), <http://www.ftc.gov/bcp/workshops/ehavioral/presentations/2mwalrath.pdf> and generally, the webscripts FTC Office of Public Affairs, "eHavioral Advertising: Tracking, Targeting, & Technology," (1 and 2 November 2007), <http://htc-01.media.globix.net/COMP008760MOD1/ftc_web/FTCindex.html#Nov1_07>.
20. Quoted in Roger Clarke "Google's gauntlets—Challenges to 'old world corps', consumers and the law," (2006) 22:4 *Computer Law & Security Report* 288–298, at p. 291, <http://www.anu.edu.au/people/Roger.Clarke/II/Gurgle0604.html>.
21. I use Google here as an example, but Google may be replaced by any search engine provider. For instance, replace Google by Microsoft, Google Search by Microsoft Live Search, and gMail by Hotmail. Most search engines are fairly similar in their business models and operations.
22. There are other important reasons why search engines use IP addresses and cookies, such as to sharpen and improve their search results (see <http://peterfleischer.blogspot.com/2007/06/did-you-mean-paris-france-or-paris.html> for an example provided by Google's Peter Fleischer) and detecting "scams" with their business model (see "Google's response to the Article 29 Working Party Opinion on Data Protection Issues Related to Search Engines, 8 September 2008," <http://www.scribd.com/doc/5625427/google-ogb-article29-response>).

visited, the user's language preference, window size, or preferences as provided by the user during the interaction. Cookies can be read by the web server that set the cookie.[23] Therefore, when a user revisits the search engine, it will know because it automatically receives the cookies it set during the previous visit. Moreover, the identifier stored in the cookie allows the web server to relate the user's current activity to whatever the server has stored about previous interactions involving the same identifier. Therefore, if a search engine stores the cookies it receives back from revisiting web browsers along with the queries submitted by these browsers, it will have a comprehensive background of the search history of this particular browser. Needless to say, the analysis on this history can be done to infer habits and interests about the user of this particular browser.

At this point, it is important to note that cookies are browser based. I use the Firefox, Safari, and Shiira web browsers on my machine during work, and the same browsers on my private account on the same machine. Each browser-user combination will have its own cookies for every site from which it receives cookies. Therefore, I will most likely have at least six cookies set by Google Search, six set by Yahoo, and so on. When I use Firefox, the search engine cannot read the content of the cookies it sent to me while I was using Safari earlier on that same day. Nor can it access the Firefox cookie on my private account during interaction from my work account, even though these two accounts reside on my Macbook.

The second method of identification involves IP addresses. IP addresses, as outlined above, identify machines. Search engines store the IPs of their users' machines along with their queries. The search history associated with particular IP addresses is therefore available to the search engine provider. In contrast to cookies, the provider can link queries submitted by different browsers and different users on the same machine on the basis of an IP address because this address will be the same in all instances. This does not make IPs more useful for the purposes of tracking individual users per se because in many cases IP addresses are (pseudo) dynamic. For instance, many internet users are assigned different IP addresses by their Internet Service Provider (ISP) on different dial-in sessions. Or in the case of broadband connections, the ISP may occasionally reassign IP addresses to prevent users from running certain software (for example, web servers). Users may also share the same IP address, for instance because their web traffic is routed through a company proxy, or they share a common internet access point (for example, a household broadband router) which makes the behaviour associated with that IP address the behaviour of multiple users. Therefore, in many cases, IPs are not suitable to identify specific individuals accurately.[24]

The two techniques can also be combined. This limits the drawbacks mentioned for the singular use of cookies or IPs. Combining cookies and IP addresses allows the server, for instance, to notice that different queries submitted by a certain IP address come from different instances of a particular browser. Because the HTTP header information received by the server with each request

23. Only machines in the originating domain and its sub-domains can read the cookies provided by those domains for obvious security reasons. However, there are workarounds to allow for cookie sharing. These require the cooperation of the issuing server. See Wayne Berry, "Sharing Cookies Across Domains," <http://www.15seconds.com/issue/971108.htm> for a detailed explanation of a way to share cookies.
24. In the case of IPs assigned during dial-in, the ISP will be able to make a connection between the IP and the customer on the basis of their logs. But within a student dorm or house where many people share the same IP, this does not work. Here the IP will point to the person who contracted with the ISP.

contains additional information, such as browser type and version and operating system type and version, more fine grained distinctions can also be made.

Do search engines engage in determining user habits beyond superficial analysis of current queries? Search engine providers are not very transparent about this.[25] What is certain is that they have the potential to do so. Search-related data, including IP addresses, cookie identifications, user identities, and search terms, are retained by search engine providers between 13 and 18 months.[26] In July and August 2007, influenced by the growing pressure from European and United States legislators, major search engine providers, including AOL, Google, Ask.com, Yahoo, and Microsoft, tumbled over each other to change their data retention regimes.[27]

As we have seen, the advertisement-serving industry and search engine providers generally do not consider cookies and IP addresses to be personally identifiable information and downplay the issues surrounding the storage of search data associated with these identifiers. Closer inspection of the data stored by these service providers, however, identifies at least two issues.

The first issue relates to the question of whether search data is indeed unlinkable to named individuals. In some instances, search data can be associated with named individuals. People frequently engage in vanity searches or self-googling queries and therefore give away information pertaining to their civil identity in the query.[28] This presents a problem even if identifying data, such as the cookie identification or the user's IP address, are replaced by a (one-way) hash code[29] or by a random number that is supposed to make the data anonymous as is eventually done by search engines. This problem was illustrated when America Online in August 2006 released pseudonymised search data relating to 650,000 of its users. User account identifications were replaced by random numbers. Journalists of the *New York Times* had little trouble revealing the identity of user 4417749 by exploiting her vanity searches which were clearly visible in this user's history.[30] This evidences that large data sets containing search data will likely reveal sufficient clues to trace back to individuals in the real world.[31] Pseudonymising the data by replacing IPs with hashes, which make identifying the user on the basis of the IP (through consulting the ISP that supplied the IP) impossible, does not therefore solve all identification issues.

25. Clarke, *supra* note 20 at p. 297, writes in this context that, "There is no evidence that the Google corporation has yet moved to bring the full power of data mining technology to bear on this rapidly growing mound of data. But that would in any case be a strategically unwise manoeuvre at this early stage."
26. See Declan McCullagh, "How search engines rate on privacy," CNET News.com (13 August 2007), <http://www.news.com/2102-1029_3-6202068.html> for an overview of how the major search engines were rated on privacy aspects.
27. Google in their response to the Article 29 Working Party Opinion on Data Protection Issues Related to Search Engines, *supra* note 22, claims, amongst other reasons, that it needs to retain these data for a long time to detect "foul" play with their rating and advertisement click-through model.
28. Soghoian, "The Problem of Anonymous Vanity Searches," *supra* note 18.
29. A one-way hash code is a function that takes a string of arbitrary length as input and deterministically produces another string with a fixed length as output. It should be extremely difficult to reverse the process. Note that anonymous in this connotation means unlinkable to a known person. More on this topic is to come later in this paper.
30. Soghoian, "The Problem of Anonymous Vanity Searches," *supra* note 18. On November 14, it was still easy to find out the real identity of AOL user 4417749, by simply entering the number in Google. In fact, for the very first results, see Michael Barbaro and Tom Zeller Jr., "The face behind AOL user 4417749," *International Herald Tribune* (15 August 2006), <http://www.iht.com/articles/2006/08/09/business/aol.php?page=1> which gives away this user's identity.
31. See also Bradley Malin, Latanya Sweeney, and Elaine Newton, "Trail Re-Identification: Learning Who You Are From Where You Have Been," in Carnegie Mellon University, Laboratory for International Data Privacy, *LIDAP-WP12* (March 2003), <http://privacy.cs.cmu.edu/dataprivacy/projects/trails/trails1.pdf>.

The second issue concerns what can be done with data inferred from search data of unnamed individuals. As we have seen, the search queries themselves reveal information about the users' interests. This can be supplemented by other information sent to the search engine automatically when the search query is submitted. The HTTP header contains data such as the user's computer and operating system (for example, Macintosh Intel Mac OS X, Windows NT 5.1) and browser type (for example, Mozilla or Internet Explorer). The IP address reveals (inaccurate) information about the geographical location of the user's machine.[32] This combined information can help the search provider to offer the user advertisements of a local Apple store when they search for "Apple bluetooth keyboard," or allow internet users in Miami to be spared advertisements for winter tires.

The analysis of search histories can be used to infer much more about an individual user. Although it may increase search precision and the relevance of advertisements presented to the individual users, practices such as knowledge discovery in databases, dataveillance, and profiling may also have adverse effects for the individual user.[33] Websites offer the possibility to completely tailor the information presented to individual users (both content and advertisements), which cannot be accomplished through traditional broadcast media, such as television. An effect of this may be that advertisements and content converge on the interests of an individual as perceived by the information provider—and by those who pay for providing the information—produce tunnel vision. Over time, this may lead to cumulative effects and self-fulfilling prophecies that further affect an individual's autonomy to make choices. Paul Schwartz has called this the "autonomy trap."[34] It also limits serendipity, which is important to spark new ideas.

The potential use of information inferred from online habits can go much further than just providing more relevant advertisements.[35] Profile data, especially if provided to third parties, may be used for social sorting and discriminatory practices, such as dynamic pricing and price discrimination. While these practices have always existed and often are perfectly within the boundaries of the freedom to enter into contracts, implementing them on a large scale was until recently prohibitively

32. See for instance sites such as IP Location Finder, <http://www.iplocationfinder.com/location.htm> and IP-Address.com, <http://www.ip-adress.com>, which provide this kind of location data on the basis of public registers such as the WHOIS database. In the author's case, these services were off by about 1 km at the time of writing this paper. The IP of the author's home computer in the Netherlands is mislocated by tens of kilometers.

33. For more on the (adverse) effects of data mining in the kind of data central to this article see, for instance, Tal Z. Zarsky, "Desperately Seeking Solutions: Using Implementation-Based Solutions For The Troubles Of Information Privacy In The Age Of Data Mining And The Internet Society," (2004) 56:1 *Maine Law Review*, 14–59, <http://law.haifa.ac.il/techlaw/papers/zarsky-maine.pdf>. For an extensive overview of knowledge discovery in databases (including data mining) and profiling see Bart Custers, *The Power of Knowledge: Ethical, Legal and Technological Aspects of Data Mining and Group Profiling in Epidemiology* (Wolf Legal Publishers, 2004). See also Roger Clarke, "Information Technology and Dataveillance," (1988) 31:5 *Communications of the ACM* 498–512, <http://portal.acm.org/citation.cfm?doid=42411.42413> about dataveillance in general.

34. Paul M. Schwartz, "Internet Privacy and the State," (2000) 32:815 *Connecticut Law Review* 821–828, <http://papers.ssrn.com/so13/papers.cfm?abstract_id=229011>.

35. Search engine results also depend on the country of origin. See for instance, Jonathan Zittrain and Benjamin Edelman, "Localized Google Search Result Exclusions: Statement of Issues and Call for Data," Harvard Law School: Berkman Center for Internet & Society (22 October 2002), <http://cyber.law.harvard.edu/filtering/google/>.

expensive.[36] The internet makes it possible to offer each individual different terms and conditions at little cost, without the user being aware of this.

Even more powerful than the collection of data about internet user preferences and reusing identifiers (such as cookies and IP addresses) are advertisement-serving companies, such as Doubleclick and Tacoda,[37] which act as intermediaries between advertisers and the media (for example, websites and publishers). They determine which advertisements are placed on a publisher's website on the basis of the data they collect about individuals' online habits and information funneled to them by the publishers. The advertisements provide the publishers with advertising revenues, which allow them to provide free content. Many of these sites make use of a limited number of advertisement servers. Because advertisement servers can recognize the user's machine (through cookies and IP addresses) and know which site the user is visiting (the request to display a banner comes from the visited site), they are able to track individual user behaviour across websites.[38] The tracking of users across websites also means that they are able to track users across different social contexts, such as work, hobby, sport, and family life. This undermines what Goffman[39] termed "audience segregation," the individual's capability to play different roles and give specific performances to specific audiences. The power to keep audiences distinct and reveal different aspects of oneself in different contexts is deemed an essential characteristic of our lives.[40]

A more detailed account of the (adverse) effects of logging online behaviour linked to IP addresses and cookies and the profiling and knowledge discovery on the basis of such data is beyond the scope of this paper.[41]

*

4. DECONSTRUCTING IDENTIFIABILITY: L-, R-, C-, AND S-IDENTIFIABILITY

WE CAN NOW RETURN TO IDENTIFIABILITY. In the introduction, I stated that the advertisement industry tries to downplay the consequences of what they do by pointing out that personal data (in the limited sense, meaning directly identifying data) is not being collected. While this is partially true, the previous section has argued that even without collecting names, numerous privacy issues are engaged.

The Data Protection Directive distinguishes between different kinds of identification, direct and indirect, and acknowledges that identification that

36. In 2000 there was a huge public outcry over Amazon's experiment with dynamic pricing, see for instance Wendy Melillo, "Amazon Price Test Nets Privacy Outcry," AllBusiness (2 October 2000), <http://www.allbusiness.com/marketing-advertising/4188108-1.html>.
37. Not surprisingly, both have been taken over by search engine providers. Google has acquired Doubleclick for $3.1 Billion, while Tacoda was bought by AOL for an undisclosed amount. See Elinor Mills, "AOL Buys ad firm Tacoda," CNET News.com (24 July 2007), <http://news.com.com/AOL+buys+ad+firm+Tacoda/2100-1024_3-6198613.html>.
38. I leave aside here the more intricate mechanisms for tracking across sites involving third-party cookies, such as webbugs, which are also known as web beacons, tracking bugs, pixel tags, 1x 1 gifs, and clear gifs. Wikipedia gives a clear account of how these function at <http://en.wikipedia.org/wiki/Web_bug>.
39. Erving Goffman, *The Presentation of Self in Everyday Life* (University of Edinburgh, 1956) pp. 41–43.
40. James Rachels,*Can Ethics Provide Answers? And Other Essays in Moral Philosophy* (Rowan & Littlefield, 1997), pp. 145–154.
41. For discussions of the risks of these practices see, for instance, Zarsky, "Desperately Seeking Solutions, *supra* note 33, and Custers, *The Power of Knowledge*, *supra* note 33. See also Tal Z. Zarsky, "Mine Your Own Business!: Making The Case For The Implications Of The Data Mining Of Personal Information In The Forum Of Public Opinion," (2002-2003) 5 *Yale Journal of Law & Technology*, pp. 2–56, <http://www.yjolt.org/old/files/20022003Issue/Zarsky.pdf>; and Greg Elmer, *Profiling Machines: Mapping the Personal Information Economy* (MIT Press, 2004).

does not result in the individual's name is identification. But at the same time one has to realize that EU data protection legislation was introduced at a time when personal data processing was different than what we are considering in this article. When the DPD provisions were drafted, data processing was done by companies and governments in face-to-face interactions with customers and citizens and by manually entering forms. The data was stored locally in (large) databases and data was exchanged on tapes and floppy disks. Computer networks were uncommon. The Directive came into effect in 1995, meaning that the early drafts were made when cookies were made of flour and butter, not bits.[42] The data protection legislation clearly shows its roots in the traditional files and folders that store patient records, customer data, government databases, and the like. One may therefore doubt whether the regulation was sufficiently prepared for what was to come.[43] Of course, relevant regulation was enacted after the Data Protection Directive including, the eCommerce Directive,[44] the Privacy and Electronic Communications Directive,[45] and the Data Retention Directive;[46] however, the foundation has not changed since 1995.

In my view, we should unravel the notions of personal data and identifiability in order to address the issues raised in the previous sections in a more comprehensive way.[47] A first step would be to clearly distinguish between two major types of identifiability instead of conflating them into a single definition. For lack of better terms, I will call them L-identifiability for Look-up identifiability,

42. Cookies were first implemented by Netscape's Lou Montulli in July 1994. See Jay P. Kesan and Rajiv C. Shah, "Deconstructing Code," (2003-2004) 6 *Yale Journal of Law & Technology* pp. 277–389, <http://www.yjolt.org/files/kesan-6-YJOLT-277.pdf> for a history of http cookies.
43. Various EU member states have or are in the process of evaluating their data protection regulation, and also the EU itself is in the process of evaluating the DPD. The results of these evaluations will give more insight as to whether the regulation is indeed fit for today's world wide web and current practices.
44. European Community, *Commission Directive 2000/31/EC of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market (Directive on electronic commerce),* <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32000L0031:EN:HTML>, [2000] *Official Journal of the European Union* L 178/1.
45. European Community, *Council Directive 2002/58/EC of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications),* <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32002L0058:EN:HTML>, [2002] *Official Journal of the European Union* L 201/37 [Privacy and Electronic Communications Directive].
46. European Community, *Council Directive 2006/24/EC of 15 March 2006 on the retention of data generated or processed in connection with the provision of publicly available electronic communications services or of public communications networks and amending Directive 2002/58/EC*, <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32006L0024:EN:HTML>, [2006] *Official Journal of the European Union* L 105/54.
47. See also Gary T Marx, "Identity and Anonymity: Some Conceptual Distinctions and Issues for Research," in Jane Caplan and John Torpey, eds., *Documenting Individual Identity: The Development of State Practices in the Modern World* (Princeton University Press, 2001) 311–327, <http://web.mit.edu/gtmarx/www/identity.html>, who, at p. 312, distinguishes seven types of identity knowledge with different degrees of identifiability: "(1) legal name; (2) locatability; (3) pseudonyms that can be linked to legal name and/or locatability [pseudo anonymity]; (4) pseudonyms that cannot be linked to other forms of identity knowledge [real anonymity]; (5) pattern knowledge; (6) social categorization; and (7) symbols of eligibility/non-eligibility." Marx's types 1, 2 and 3 are L-identifiers in my terminology; 4, 5, 6 and 7 are R-identifiers; and 6 is a C-type identifier. Another relevant distinction in this respect pertains to authentication factors, consisting of pieces of information used to authenticate or verify an individual's identity: something the user has (e.g. key, card, document), something the user knows (e.g. pincode, password), or something the user is or does (e.g. photograph, fingerprint). See, for instance, Bruce Schneier, *Secrets and Lies: Digital Security in a Networked World* (John Wiley, 2000). Authentication factors usually come into play after an individual is identified.

and R-identifiability for Recognition identifiability.[48] [49]

4.1. L-identifiability

ALL FOUR TYPES OF IDENTIFIERS ALLOW INDIVIDUALS to be identified. The essential characteristic of an L-identifier is that there is a register, directory, or table that provides the connection between the identifier and a named individual—hence I call this kind of identifiability *look-up identifiability*. Names, telephone numbers, passport numbers, social security numbers, and IP addresses are examples of L-identifiers. Because there is a connection between the L-identifier and a named individual (civil identity), L-identifiers can be used beyond identification. Someone who has access to an L-identifier can discover to whom in the real world the identifier belongs and can therefore address this individual outside of the context in which the identifier is used.

Suppose, for instance, that a video rental shop requires their customers to use their social security number as their usernames; in that case, having access only to the list of usernames would be sufficient to create a list of the video rental shop's customers. A competing video shop who gains access to this list could then target these individuals with special offers to join their service. Or less innocently, access to the names of the customers may trigger further investigation into their habits.[50]

L-identifiability is not a zero-one matter. Discovering to whom a certain L-identifier belongs may range from relatively easy, as in the case of a telephone number, to extremely difficult, as in the case of a passport number. Also, the effort required differs from one individual to the next. Finding out my name on the basis of my passport number is easy for a civil servant working at the registrar in the Netherlands, whereas this task would be challenging for most readers of this paper.

Some L-identifiers identify more precisely or uniquely than others. Consider the difference between a driver's license number and an IP address. The driver's license number is uniquely associated to a single individual. IP addresses are not always uniquely associated with an individual, and not even to a single machine. Many IP addresses allow the identification of concrete internet users, or a limited set of users (e.g. a household) because the internet service providers can often make the connection between the IP address and a natural person (the subscriber). There are also exceptions as we have seen, such as internet cafes with shared and dynamic IP addresses without users having to register themselves. Sometimes additional data is required to make the connection. In the case of dynamically assigned IP addresses, for instance, it is necessary to

48. A recent proposal submitted to the Federal Trade Commission for the "eHavioural Advertising" workshop by the Center for Democracy and Technology, Consumer Action, the Consumer Federation of America, the Electronic Frontier Foundation and other institutes, is close to making a similar distinction in their proposal for a definition of Personally Identifiable Information. See Ari Schwartz et al., "Consumer Rights and Protections in the Behavioral Advertising Sector" (Center for Democracy and Technology, 2007), <http://www.cdt.org/headlines/1057> .

49. One could also use *findability* or *locatability* for L-identifiability, *recognizability* for R-identifiability, and *classifiability* for C-identifiability as synonyms, but this obscures their relatedness.

50. That this can indeed have serious consequences as illustrated in the famous disclosure of US Supreme Court nominee Robert Bork's video rental records in a newspaper in 1988, which also led to the enactment of the 1988 Video Privacy Protection Act in the US. See Electronic Privacy Information Centre, "The Video Privacy Protection Act," <http://epic.org/privacy/vppa> (6 August 2002). See, for instance, Daniel J. Solove, *The Digital Person: Technology and Privacy in the Information Age* (New York University Press, 2004), <http://ssrn.com/abstract=609721> at p. 69.

provide an exact date and time in order for the ISP to determine to whom the IP address was assigned at that moment. But on the whole I would argue that IP addresses are indeed linkable to individuals, a point of view also adopted by the Article 29 Working Party.

In the parlance of the Data Protection Directive, most L-identifiers belong to the category of indirect identifiers as specified in article 2 of the DPD because additional data is required to get to the individual in the real world (through her name/address). The L-identifier name (and direct ancillary data) is of course the exception. Names are direct identifiers according to article 2 of the DPD.[51]

4.2. R-identifiability

R-IDENTIFIERS ARE IDENTIFIERS THAT ALLOW AN individual to be *recognized* without being able to associate the identifier with a named individual, hence I call this kind of identifiability *recognition* identifiability.[52] R-identifiers require the presence or activity of the individual. The individual is recognized because she presents an identifier, token or feature set (e.g. description of physical appearance), known or recognizable as valid by the recipient, to the entity performing the identification. R-identifiers derive their meaning from the fact that the recipient accepts the identifier as a valid identifier. The bearer or presenter of the identifier is identified by virtue of the presentation of the identifier.

The realm of an R-identifier is that of the context in which it was created and there are no ways to tread outside this realm, certainly not in the real world. R-identifiers are therefore more confined in their operational scope than L-identifiers.

R-identifiers are fairly common and have existed for a long time.[53] Tokens are credentials that establish a right to claim of a certain set of attributes.[54] They allow the recipient to recognize the bearer as being someone, or something, being entitled to something or as having some attribute or property. Cloak room tokens and bearer checks[55] are common examples. These certificates are used in the context of authentication for a particular claim. Authentication answers the questions "Who are you?" and "How do I know I can trust you?"[56] In the case of the cloak room token, the token identifies the bearer as the purported owner of said coat, and presenting a genuine looking token is supposed to convey trust that the reclaim of the coat is valid. R-type tokens allow the recipient to identify the presenter as entitled to something, without disclosing the bearer's

51. Data Protection Directive, *supra* note 5 at art. 2.
52. If the R-identity was issued by the entity making the identification, then R-identifiers allow for the verification of the individual's identity.
53. For instance, Wikipedia, "Cheque," <http://en.wikipedia.org/wiki/Cheque#_note-Vallely> mentions that "[i]n the 9th century, a Muslim businessman could cash an early form of the cheque in China drawn on sources in Baghdad, a tradition that was significantly strengthened in the 13th and 14th centuries, during the Mongol Empire. Indeed, fragments found in the Cairo Geniza indicate that in the 12th century cheques remarkably similar to our own were in use, only smaller to save costs on the paper. They contain a sum to be paid and then the order 'May so and so pay the bearer such and such an amount.' The date and name of the issuer are also apparent."
54. Philip J. Windley, *Digital Identity: Unmasking Identity Management Architecture (IMA)* (O'Reilly, 2005) at p. 50.
55. A bearer check is payable to anyone who is in possession of the document.
56. Windley, *Digital Identity*, *supra* note 54 at p. 50.

civil identity.[57]

On the internet R-identifiers are common. Cookies are examples of R-type identity credentials, as are certain usernames and raffle or sweepstake tokens. They tie transactions together that are otherwise difficult to connect.[58] Their popularity derives from this characteristic. In many situations there is no need whatsoever to go beyond being able to reconnect individuals to previous transactions. R-identifiers provide just that. They enable personalization of the "experience" and allow service providers to build and use files about their users. In many cases the issuer of the R-identifiers has no interest in the individual's name or civil identity, and consciously or unconsciously has decided not to ask the user to provide personal data and chosen to use an R-identifier instead of an L-identifier.[59].

*

5. THE RELATION BETWEEN L-IDENTIFIERS AND R-IDENTIFIERS

THE DISTINCTION BETWEEN L-IDENTIFIERS and R-identifiers comes to light when we consider the two prevalent identifiers on the internet discussed in the previous section: IP addresses and cookies. They are used in similar ways. Cookies and IP addresses are the keys to files maintained by service providers about their users. When users visit a service provider's website, they automatically present these keys to the web server allowing the web server to retrieve their file. Both kinds of identifiers also likely qualify as identifying data in light of the Data Protection Directive, although this is not entirely certain and awaits pending research by the Article 29 Working Party. So from this perspective it would appear that cookies and IP addresses are very similar. When approaching them from the distinction introduced, there appears to be a clear difference. In the case of IP addresses there is a serious chance that the civil identity of the user of the IP address can be revealed. Therefore, IP addresses belong to the realm of L-identifiability. Determining the civil identity of a user on the basis of a cookie is impossible.[60] Cookies are just (random) tokens issued by a website to be recognized later as issued by the same website. Cookies therefore belong to the realm of R-identifiability.

R-identifiers can be transformed into L-identifiers by centrally storing them and associating personal data with them. Therefore some identifiers can be used as either L- or R-identifiers. This is the case with biometric data, such as fingerprints, or retinal data (iris scans), and data such as DNA samples. If the

57. Sometimes, the credential does contain such information, but that is either the result of the multiple purposes the token serves (e.g. my ticket for a dance performance at the National Arts Centre in Ottawa contained my name, because the ticket also served as the receipt for my credit card payment), security requirements (e.g. driver's license which certifies that the holder is entitled to drive a car, but which also has a photo to limit fraud), or plain ignorance of the issuer (e.g. my biometric trusted traveler pass (PRIVIUM, *infra* note 62) does contain my name in print. Because the card is only used by machines that read the data on the card's chip and verify my iris scan with the template on the chip, my printed name is irrelevant.).
58. Windley, *Digital Identity*, *supra* note 54 at p. 51.
59. An example of a deliberate choice to use an R-identifier instead of an L-identifier is the following. Within the PRIME project (Privacy and Identity Management for Europe (<http://prime-project.eu>) we have conducted an online survey using questionnaires that featured a sweepstake for the participants. Given the survey's topic, privacy, and our purpose, collecting anonymous responses, we clearly did not want to collect personal data and therefore decided to issue randomized tokens to the participants who completed the survey. The contestants could check their eligibility to a prize by entering their token. Only the winners had to disclose their address in order to receive their prize. Also this last step could have been done anonymously, but that would have made the process cumbersome for both winners and researchers.
60. Unless, of course, the user registered herself on the service provider's website when the cookie was placed on her machine.

data (or the templates derived from the raw data) is stored in central databases[61] together with the names of their bearers, these are clearly L-identifiers. A particular biometric sample can be compared with the data in the database to reveal the name (and other data) of the bearer (identification). These samples can equally be used as R-identifiers in which case only verification of the bearer against the sample can be conducted. This requires local storage of the biometric data (or template) on something under the control of the individual, such as a smart card. This is the case in certain trusted passenger schemes, including Schiphol Airport's PRIVIUM system.[62] The biometric sample in this case functions as an R-identifier allowing a machine to recognize the holder of the card as being the person to which this card was issued (verification). Regarding another biometric, fingerprints, the Dutch government has decided to use them as L-identifiers. As of 21 september 2009 four fingerprints of each applicant of a Dutch passport or identity card will be stored not only on the chips embedded in these photo-IDs, but also in a central database. The government here has moved beyond the EU prescribed obligation to incorporate fingerprints in the passport.[63]

5.1. C-identifiability

THE THIRD TYPE OF IDENTIFIABILITY IS C-IDENTIFIABILITY, or Classification identifiability. In the case of C-identifiability, there is a set of preexisting group profiles or categories,[64] and individuals are classified as belonging to one or more of these categories on the basis of their interaction with a particular website. Users are therefore identified as members of a particular group or category. In the case of C-identifiability the purpose of identification is not so much to recognize the individual as an individual, but rather to classify the individual as an instance of a class the website knows about. The classification will bring the service provider's knowledge about the class to bear on the individual: certain beliefs and practices are attributed to the individual (ascription[65]). A hypothetical example is the following. An online bookstore, let's call it Wolga.com, distinguishes chick lit readers, cruel crime readers, real crime readers, and romantic readers, among other categories. On the basis of the browsing behaviour of a certain visitor, the website's classification algorithm may decide that the visitor is a chick lit fan and consequently present recommendations relating to chick lit. This process of ascribing certain attributes to an individual can, of course, take more serious forms. This is what knowledge discovery in databases is about—finding categories and clusters of related data and being able to associate (meaningful)

61. This is increasingly the case for DNA. See, for instance Home Office, "National DNA Database," <http://www.homeoffice.gov.uk/science-research/using-science/dna-database>.
62. PRIVIUM, <http://www.schiphol.nl/privium/privium.jsp>.
63. This follows from the new Dutch Passport legislation entering into force on 21 September 2009 <http://www.paspoortinformatie.nl/content.jsp?objectid=4495>. The move to create a central biometric database for Dutch fingerprints is made by the Dutch government in an attempt to fight look-alike identity fraud, aid law enforcement and aid identification of disaster victims. Note that once biometric data is transformed from an R-identifier into an L-identifier, there is no way back as long as the register exists because there is always the option of comparing the sample to the data in the register. This is one of the reasons to be particularly careful with biometric data.
64. These categories may be derived from data mining techniques as part of Knowledge Discovery in Databases (KDD). In data mining, knowledge discovery techniques such as regression analysis, cluster analysis and classification are used. See Custers, *The Power of Knowledge*, *supra* note 33; Zarsky, "Mine Your Own Business," *supra* note 41. Some techniques are hypothesis driven, whereas others merely look for statistical patterns.
65. See Custers, *The Power of Knowledge*, *supra* note 33 at p. 58.

labels with them, which can subsequently be associated with individuals or groups, which are then believed to have certain beliefs or properties.[66]

C-identifiability is related to R-identifiability in the sense that in both cases the real world identity of the individual is irrelevant. However, in the case of R-identifiability, the identifier is issued by the service provider to the individual (e.g. a cookie).[67] In the case of C-identifiability, the service provider distinguishes a set of group profiles and associates a set of attributes or rules with each of these profiles. Their labels are their C-identifiers. C-identifiers live in the service provider's realm, whereas the R-identifier is issued to the user. For instance, a rule associated with the chick lit profile may be something like: "activate when a user conducts multiple searches for authors belonging to a predefined group of chick lit writers, or clicks on any of the writers on this list." The users, in their interaction with the website, will trigger one or more of these rules by virtue of their online behaviour and the attributes thereby displayed. A chick lit reader will perform the kind of behaviour displayed in the rule, and therefore be labelled as an instance of the class denoted by the C-identifier.

In the case of R-identifiability, the identifier is a token that allows the issuer to recognize the individual. Usually, there will be a file on this particular user that will be brought into play following the identification. This file may be constructed from scratch on the basis of the interaction between the user and website. In the case of C-identifiability there always is pre-existing knowledge about the type of user that, on the one hand, allows the association of the user with a specific class, and, on the other hand, contains basic data about this user in a way that resembles the record constructed in the case of R-identifiability. So the typical procedure in the case of a C-identifier will be: recognition of the user as an instance of a class, issuing an R-identifier for future use, establishing an R-type record about the user, and associating the C-type profile data to this record.

5.2. S-identifiability

THE FINAL TYPE OF IDENTIFIABILITY IS S-IDENTIFIABILITY, or session identifiability. S-identifiers are identifiers that allow a web server to track a user during a particular interaction and their lifetime typically is a single "session." An ecommerce site may, for instance, place an identifying cookie on the user's machine when she enters the online store in order to track the user throughout the shopping experience. The cookie here allows the server's software to pick out the correct shopping cart when the user moves between shopping and browsing through the shop. In most cases, there are different technical solutions to maintain track of the user throughout the site, but cookies are a simple and straightforward way to solve the problem of the statelessness of the web. HTTP is a stateless protocol—every page request to a web server looks like a different session, which makes it impossible for a website to run a shopping cart. Cookies were designed to solve this problem, by allowing the web server to keep track

66. See Zarsky, "Mine Your Own Business," *supra* note 41 and Custers, *The Power of Knowledge*, *supra* note 33.
67. Possible exceptions involve instances where some mutually known feature set is used as the identifier, for example when a rose is used as an identifier in a blind date.

of page requests belonging to a single session.[68]

S- and C-identifiers represent different dimensions of identification than L- and R-identifiers and they serve different purposes. L- and especially R-identifiers embody a temporal dimension; they are relevant for the future and allow the service provider to recognize returning individuals. S- and C-identifiers serve their goal in the session in which they are created (S-identifier) or invoked (C-identifier) and are even useful to service providers if their lifespan is confined to this single interaction. If a persistent connection between the individual and the data on the server is required, their role will be taken over by an R-identifier that will be issued by the service provider during the session.

In everyday life, all four types of identifiers will be used in online interactions. Although it is possible to implement a web shop without identifiers, this is rarely the case in practice. If we look at real websites, such as Amazon.com, we will see all four types of identifiers in action in the case of registered customers. Amazon will place an R-identifying cookie on the user's machine to facilitate recognizing the user as a returning Amazon visitor. When a registered user logs in, one of the cookies Amazon has placed on the user's machine will act as a pointer to Amazon's records of the user. These records will contain one or more L-identifiers (name, address, etc.) of the user. When the user goes shopping, one of the (temporary) cookies will serve as a session identifier to keep the proper shopping cart associated to the user. And finally, Amazon will probably use their group profiles and other mechanisms to try to figure out what the user's preferences are, including by watching out for C-identifiers created by the user as a result of her activities in the store, which can be associated with the proper group profiles by Amazon behind the scenes.

*

6. USING THE DISTINCTIONS

THE REASON THE DISTINCTION BETWEEN THE FOUR TYPES of identifiers is useful is that it helps with analyzing the issues and devising proper solutions. The Data Protection Directive in its current form treats all kinds of collection of personal data alike. When data can be qualified as personal data, as defined in article 2 of the Directive,[69] the Directive applies and with it all the obligations on data controllers and processors and the rights of the data subjects come into play. From thereon there are few distinctions in obligations and rights.

In practice some obligations and rights are spurious and lead to objections by the industry that, considering the distinction introduced in this paper, make sense. For instance, article 5(3) of the Privacy and Electronic Communications Directive requires "clear and comprehensive information […] about the purposes of the processing […]" of cookies and requires service providers to offer the user the "right to refuse such processing by the data controller."[70] The provision contains an exception for "any technical storage or access for the sole purpose of

68. See Kesan and Shah, "Deconstructing Code," *supra* note 42, for a history of how Netscape's Persistent Client State HTTP Cookies solved the statelessness problem faced by the Netscape Enterprise Server Division.
69. Data Protection Directive, *supra* note 5 at art. 2(a).
70. Privacy and Electronic Communications Directive, *supra* note 45 at art. 5(3); Sylvia Mercado Kierkegaard, "How the *Cookies* (Almost) Crumbled: Privacy & Lobbyism," (2005) 21:4 *Computer Law & Security Report* 310–322 at p. 320.

carrying out or facilitating the transmission of a communication over an electronic communications network, or as strictly necessary […] to provide an information society service explicitly requested by the subscriber or user."[71] While the second part of this provision seems to "pull the rug" from under the first part,[72] the scope of the exception is not entirely clear and in any case awkward. S-type identifiers certainly are covered by the exception, but what about R-identifiers whose sole purpose is to activate user preferences or user settings on return to a site? If all R-identifiers fall under the exception, then indeed the rug is pulled from under article 5(3). If all R-identifiers fall under the main rule, then one may question why innocent R-identifiers set for the purposes of restoring settings and preferences have to be preceded by detailed information and explicit options for opt-out.[73]

Making the distinction between L-, R-, and C-identifiability explicit makes it easier to specify separate regimes for the collection and use of data that somehow relate to individuals in online interactions. L-, R-, and C- identifiability raise different concerns and different regulatory regimes may therefore be appropriate. In the remainder of this paper, I will provide some glimpses on what this could mean. Grasping the full complexity is beyond this paper and requires much more study.

6.1. L-identifiability

L-IDENTIFIERS MAKE IT POSSIBLE TO OBTAIN DATA directly relating to named individuals in the real world. This facilitates the tracking and addressing of individuals outside the scope of the interaction or relation in which the L-identifiers play a part. Data controllers and third parties can therefore use L-identifiers to initiate new interactions and relations or enter the intimate and private sphere of the individual. A car insurance telemarketer could, for instance, use my name, which I may have disclosed on the website of a car dealer, and my approximate location conveyed by my IP address to try to find out my telephone number and call me to offer me car insurance packages. The telemarketer can therefore contact me only if they have relatively harmless information, such as the names and IP addresses of people visiting the car dealer's website. Still, particularly if I don't have an existing relation with the car insurer, I may not approve of the privacy breach caused by the phone call. Nor am I particularly fond of people knocking on my door as a result of them finding out my address on the basis of my IP address.

L-identifiers can be used behind the back of the concerned and without their knowledge, which is, given the nature of possible privacy breaches, undesirable. Informed consent of the data subject, regarding the collection and use of data directly relating to the named individual, seems an appropriate mechanism to mitigate harms. Requiring websites to provide information about the L-identifiers and other personal data they collect and how they are used, in addition to offering users ways to opt-in (or at least opt-out) of such uses, therefore makes sense. Having the opportunity to state my preferences regarding (third party) use of the data could spare me phone calls by unfamiliar third parties offering me services.

71. Privacy and Electronic Communications Directive, *supra* note 45 at art. 5(3); Kierkegaard, "How the *Cookies* (Almost) Crumbled," *supra* note 70 at p. 320.
72. Kierkegaard, "How the *Cookies* (Almost) Crumbled," *supra* note 70 at p. 320.
73. This can also be done within the user's browser, even though this is inconvenient.

Regarding the rights of data subjects, individuals clearly have a stake in the correctness of the information pertaining to them because the data may be used not only in decisions about them in the context of relations they have entered into themselves, but also in decisions outside the realms in which they are directly involved. Hence, providing individuals the right to inspect the data associated to their L-identifier[74] and the right to have the data corrected also seems reasonable.

6.2. R-identifiability

IN ORDER TO FUNCTION, R-IDENTIFIERS REQUIRE the presence or activity of the individual to whom they pertain. The individual is recognized when their token is presented to the service provider, or when the individual's behavior allows for their recognition, for instance through the queries they submit or the clickstream they produce. The operational scope of R-identifiers is therefore more limited than L-identifiers. Their realm is that of the context in which they were created and there is no way to tread outside this realm, and certainly not in the real world.[75]

Is consent for creating, storing and using R-identifiers a useful concept? R-identifiers do relate to individuals and are used in ways that affect these individuals, but in many of their applications consent is fairly impractical and unnecessary. Cookies, for instance, provide a convenient mechanism to recognize returning users which may facilitate tailoring the interaction with the user. They can be used to store preferences or provide a link to user preferences on the service provider's website.

Cookies over time have become almost indispensable. Although it is possible to configure one's browser to (selectively) block cookies, this largely undermines the utility of the internet. Although the industry itself has created this situation,[76] it has a point in stating that: "Without cookies, the Internet would be slower, the electronic marketplace cumbersome and the entire online experience frustrating."[77]

Many cookies pose no privacy threats at all—think of the cookies that store user preferences or are the ones that maintain a shopping session. Requiring consent (opt-in) to store and use these cookies for each individual cookie would be placing too great and an unnecessary burden on the user.[78] Implementing a strong

74. I take the L-identifier here to be the minimal set of attributes required to unambiguously identify the individual, such as name and date of birth. Often the individual records maintained by the service provider include other attributes as well, such as transaction history, payment data and even contact information. The L-identifier is the key to this record.
75. Not directly that is. Things of course change when different types of identifiers can be linked. That is precisely what Google aims at achieving with their acquisition of Doubleclick, <http://www.doubleclick.com>. This kind of linking of identifiers is also what sparks concern in the privacy advocacy world.
76. The situation could have been different if the statelessness problem of the HTTP protocol would have been resolved in a different way. Cookies were hastily introduced by Netscape as a fix to the problem: "This pace left cookies as a technological kludge put together overnight." See Kesan and Shah, "Deconstructing Code," *supra* note 42 at p. 300. The Internet Engineering Task Force (IETF) drafted a standard for state management on the internet, as a response to Netscape's cookies, based on a technology different from cookies that was more sensitive to privacy. Needless to say, this standard did not make it.
77. See Emily T Hackett, "Cookie Policy," (13 December 2003) *Internet Alliance,* <http://www.internetalliance.org/pdf/cookie-policy.pdf>.
78. There are no technical means on the browser level to make a distinction between "harmless" and "harmful" cookies. The browser can be instructed how to handle "first party" cookies (that come from sites where you navigate) versus "third party" cookies (that come from other sites, such as those affiliated with the "first party" website), but this is not the same.

opt-in regime for R-identifiers[79] is throwing out the baby with the bath water.[80]

Instead, I would argue that a distinction between cookies that only facilitate interaction (e.g. user preferences, language) versus cookies that function as R-identifiers (to access and manage records about individuals on the websites of service providers) should be made possible on the technical level to allow web browsers to handle the two types differently.[81]

Instead of condemning all cookies, we should assess and handle the real issues surrounding R-identifiers. A prominent issue is the *construction* and especially *use* of profiles on the basis of which activities such as behavioral targeting and social sorting are carried out. These practices are very opaque at present. Users are largely unaware that profiles about them are being constructed, that behavioral targeting occurs and that profiles are used for making decisions about them.[82] The lack of transparency may cause internet users to distrust service providers, which in turn may lead to the alienation of internet users from industry and service providers.[83]

Profiling should not be addressed by simply placing a ban or limit on the collection and use of (personal) data. Privacy is not an absolute right, but one that has to be weighed against other interests. The European Data Protection Directive tries to strike a balance between the free flow of information[84] and the privacy interests of the individual. The free flow of information is even stronger in North America. This means that the collection of (personal) data is not forbidden per se.

The Data Protection Directive merely tries to capture a reasonable balance by defining the conditions under which personal data may be collected and processed. According to the DPD, personal data may be collected only for "specified, explicit and legitimate purposes and [may] not [be] further processed in a way incompatible with those purposes" (finality principle).[85] The data should be "adequate, relevant and not excessive in relation to the purposes for which they are collected and/or further processed" (data minimization principle).[86] Data should be "accurate and, where necessary, kept up to date."[87] Personal data should not be "kept in a form which permits identification of data subjects for

79. As has been argued when the Privacy and Electronic Communications Directive, *supra* note 45, was being drafted. See Kierkegaard, "How the *Cookies* (Almost) Crumbled," *supra* note 70.
80. That leaves unaddressed the question whether mandatory opt-out options should exist. I see no principled obstacles to this kind of safeguard under the control of the individual.
81. Basically, this calls for distinguishing types of cookies in the HTTP cookie protocol. Incorporating an attribute that signifies the cookie function in the cookie format allows web browsers to be instructed to accept certain types without involving the user. For certain other types of cookies, policy rules can be used to allow the browser to handle these to a lesser or fuller extent automatically without consulting the user.
82. In 2000, the Federal Trade Commission's report on Online Profiling cited a Business Week/Harris Poll which reported that only 40% of their respondents had heard of cookies, and of those 75% had a basic understanding of what they are. See "Business Week/Harris Poll: A Growing Threat," (20 March 2000) *Business Week*, <www.businessweek.com/2000/00_12/b3673010.htm>. See also George R. Milne, Andrew J. Rohm, and Shalini Bahl, "Consumers' Protection of Online Privacy and Identity," (2004) 38:2 *Journal of Consumer Affairs* 217–232.
83. See for instance, the Ponemon data presented at the Federal Trade Commission's Town Hall on eHavioral Advertising, Larry Ponemon, "FTC Presentation on Cookies & Consumer Permissions," (1 November 2007) *Federal Trade Commission*, <http://www.ftc.gov/bcp/workshops/ehavioral/presentations/3lponemon.pdf>. See also Joseph Turow, Lauren Feldman, and Kimberly Meltzer, "Open to Exploitation: American Shoppers Online and Offline," (1 June 2005) *A Report from the Annenberg Public Policy Center of the University of Pennsylvania*, <http://www.annenbergpublicpolicycenter.org/Downloads/Information_And_Society/Turow_APPC_Report_WEB_FINAL.pdf>.
84. The EU is after all formed to drive forward its member states' economies. See also the first preambles of the Data Protection Directive, *supra* note 5.
85. Data Protection Directive, *supra* note 5 at art. 6(b).
86. Data Protection Directive, *supra* note 5 at art. 6(c).
87. Data Protection Directive, *supra* note 5 at art. 6(d).

longer than is necessary for the purposes for which the data were collected or for which they are further processed."[88] The processing of personal data should be carried out in a fair and lawful way with respect to the data subjects (principle of fair and lawful processing).[89] Personal data processing has to be legitimate, either by the data subject's unambiguous consent, by a legal obligation, by contractual agreements or by other reasons listed in Article 7.[90]

What these conditions, which are rooted in the Code of Fair Information Practices[91] and the OECD Privacy Guidelines,[92] aim at is the decent treatment of people in society. Common decency (fair treatment) is therefore a core value of data protection.[93] Fair treatment in the online context implies that people know that data about them are collected as well as what data are collected and for what purposes these are used, irrespective of whether the data are personal data within the current definitions of the regulation. The intention of the data protection regulation goes beyond this. The position, as taken by the industry, that R-identifiers do not directly identify individuals and therefore require no special attention is untenable in my view. R-identifiers are being used as pointers to records about individuals and these in turn are used to make judgments about the individuals. This, in my view, warrants treating them as such. The conduct of service providers should be transparent and in line with essential elements in the data protection principles, such as purpose specification and purpose limitation. This means service providers should clearly specify their use of R-identifiers. Furthermore, their actual use of R-identifiers and the associated data should adhere to their stated purposes.

Beyond this, users and their concerns should be taken seriously, which means that they should have choices to opt-out of the use of data associated with R-identifiers for certain uses. The proposal of the Center for Technology and Democracy and others presented at the Federal Trade Commission's Town Hall on eHavioral targeting to install a (national) "Do Not Track List"[94] is a step along this road.

In the European context it would be conceivable that such measures are implemented in regulation.[95] In practical terms an obligation for service providers could be introduced to give their users more control over which kinds of data

88. Data Protection Directive, *supra* note 5 at art. 6(e).
89. Data Protection Directive, *supra* note 5 at art. 6(a).
90. Data Protection Directive, *supra* note 5 at art. 7.
91. See US Department of Health, Education and Welfare, Secretary's Advisory Committee on Automated Personal Data Systems, Records, computers, and the Rights of Citizens viii (1973), "The Code of Fair Information Practices," available at <http://www.epic.org/privacy/consumer/code_fair_info.html>.
92. See Organization for Economic Co-Operation and Development, *OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data*, (23 September 1980), <http://www.oecd.org/document/18/0,2340,en_2649_34255_1815186_1_1_1_1,00.html>.
93. See on this point of view, for instance, Bert-Jaap Koops, "Some Reflections on Profiling, Power Shifts, and Protection Paradigms," in Mireille Hildebrandt and Serge Gutwirth, eds., *Profiling the European Citizen: Cross-Disciplinary Perspectives* (Springer, 2008). See also Jeffery L. Johnson, "Privacy and the Judgment of Others," (1989) 23:2 *The Journal of Value Inquiry* 157–168; Gary T. Marx, "What's in a Concept? Some Reflections on the Complications and Complexities of Personal Information and Anonymity," (2006) 3:1 *University of Ottawa Law & Technology Journal* 1–34, <http://www.uoltj.ca/articles/vol3.1/2006.3.1.uoltj.Marx.1-34.pdf>.
94. See Schwartz et al., "Consumer Rights and Protections," *supra* note 48.
95. Whether to impose such obligations on the data collectors to provide meaningful choice to their users or whether the industry should be given the opportunity to self regulate is an interesting question that always spurs debate. Not surprisingly, the industry proclaims that self regulation will do the trick (see for instance Kierkegaard, "How the *Cookies* (Almost) Crumbled," *supra* note 70)**,** while privacy advocates call for regulation (see for instance the joint declaration by the Center for Technology and Democracy and others, Schwartz et al., "Consumer Rights and Protections," *supra* note 48).

(e.g. behavioral, geographical, temporal) associated with their R-identifiers may be used for specified purposes. Such a provision would require a comprehensive distinction to be made in types of data and purposes of data collection and use. Specifying both data types and purposes would allow users and industry to meet half way. A user could, for instance, specify that she does not object to the use of geographical data because this warrants her from being targeted with snow tire ads when she in fact lives in Miami, whereas at the same time she may specify that she will not permit the use of her clickstream data to profile her (for instance as a soccer mom). But are users actually capable of weighing the benefits versus the detriments of such data collection? Should the legislator play a role here for the good of society?

Our concerns about profiles should not stop here. The creation of profiles on the basis of individuals' online habits is one thing, while the use of these profiles to make decisions about these individuals is another. Discriminatory practices and unfair treatment of individuals especially come into play in the application of profiles.[96] This brings us to harms resulting from profile application. Should regulation pay more attention to redressing wrongs?[97] For this to work, the individual concerned would have to take the initiative in the process. This raises interesting issues.

In order to detect unjust treatment, the individual first has to become aware that a (potentially) unjust decision has been taken about him. Since these decisions range from showing a specific advertisement, to withholding information, or barring a service, detecting these potentially unjust treatments is far from trivial. One way of assisting individuals in this task would be to visually signal information on the screen as resulting from R-identifier associated data processing. So, for instance, ads could contain an indicator (say a colored dot) that signals that the ad was placed on the basis of some R-identifier. Clicking on the dot could then reveal how the ad got there by revealing in a comprehensive way which R-identifier was used and which entity decided to place the ad and why.[98]

This brings us to the issue of assessing the decision itself. For advertisements, just signaling the fact that they are consciously placed will be sufficient for most users, but if services are denied, merely signaling an R-identifier based decision is insufficient. In order to assess the decisions in these cases, the individual would need to have access to the data that was used to reach the decision, as well as to the logic applied to the data.[99] Should either data or logic be faulty, this would be a cause for action. Apart from the fact that it will probably not be completely trivial for ordinary citizens to understand how the conclusions are derived from the logic and data, the industry would likely not be jumping with enthusiasm to provide data and logic because these are central to their business. Correcting incorrect profile data also raises an interesting privacy dilemma: in

96. Discriminatory practices do not only arise in the application of profiles. Since decisions about what data to collect occur during online interactions, biases are also built into the creation of online profiles.
97. As, for instance, argued by Koops, "Some Reflections on Profiling," *supra* note 93.
98. This is similar to the kind of visual indicators regarding online privacy policies used in Janice Tsai, Serge Egelman, Lorrie Cranor, and Alessandro Acquisti, "The Effect of Online Privacy Information on Purchasing Behavior: An Experimental Study," (7 June 2007) *Workshop on the Economics of Information Security*, <http://weis2007.econinfosec.org/papers/57.pdf>. Ebay is testing such a system, called AdChoice, for their offsite ads. See <http://blogs.mediapost.com/behavioral_insider/?p=187>.
99. See Mireille Hildebrandt and Serge Gutwirth, eds., *Profiling the European Citizen: Cross-Disciplinary Perspectives* (Springer, 2008) at chaps. 14 and 15.

order to have incorrect profile data corrected, the individual will have to disclose data that was undisclosed up to their intervention.[100]

6.3. C-identifiability

THE PROBLEMS SKETCHED IN THE previous section apply equally to C-identifiers because they are a subcategory of R-identifiers. C-identifiers have an additional feature that requires further exploration. The website using C-identifiers will at one point decide to classify a user as belonging to a particular class. This involves making a decision in which preexisting knowledge is brought to bear and which may have serious consequences for the individual. It is worthwhile to consider whether and to what extent such decision making should be transparent to the user.[101]

What this limited inspection of R- and C-identifiability reveals is that the issues surrounding profiling boil down to the question of how we can prevent unfair practices, unfair judgments, and other adverse effects of stereotyping. This has proved very difficult, if not impossible, in the offline world. I doubt whether we can do any better in the online world. Offering more transparency could provide a starting point though.

*

7. FROM L-IDENTIFIERS TO R-IDENTIFIERS

THERE IS A TENDENCY IN THE ONLINE WORLD to collect L-type identifiers.[102] Publishers and service providers collect names, addresses and phone numbers to perform their contractual obligations, but also to address their clients in case of contractual default. Personal data in this sense helps to build online trust. But in cases where these reasons are absent, all too often enterprises still resort to collect L-identifiers, often because they are unaware of the potential negative effects this may have or of the alternatives that do exist. In other words, often privacy risks are not introduced intentionally, but rather result from ignorance.

One example where collecting L-identifiers is completely unnecessary is access logs maintained by web servers. Website access is logged on the basis of the IP address of the computer that was used for visiting the site. Until recently, the use of IP addresses for this purpose has not attracted much attention. Nowadays, service providers go to great lengths to argue that what they do does not pose privacy concerns such as those central to this paper. Here is an example. Apple's website[103] states:

> As is true of most websites, we gather certain information automatically and store it in log files. This information includes Internet Protocol (IP) addresses, browser type, Internet Service Provider (ISP), referring/exit pages, operating system, date/time stamp, and clickstream data. We use this information, *which*

100. See Custers, *The Power of Knowledge*, *supra* note 33 at p. 157. An important question is what is 'faulty' and correcting 'incorrect profile data'? Is the information inaccurate? Are the objectives of the data controller unfair to the consumer or to society at large?
101. People make these kinds of (value) judgments all the time and most of the time they do not inform the subject of their assessment.
102. See for instance Jim Harper, *Identity Crisis: How Identification is Overused and Misunderstood* (Cato Institute, 2006).
103. This is merely an example. I could have picked any company's website.

> *does not identify individual users*, to analyze trends, to administer the site, to track users' movements around the site and to gather demographic information about our user base as a whole.[104]

If Apple is not interested in L-identifiability for this purpose, then why don't they use R-identifiers that cannot be linked to named individuals to accomplish the same ends instead? To make it concrete consider the following alternative procedure. When a user first enters the Apple site, the web server obtains the user's location from their IP address in order to provide region specific information. Next it creates a cookie to be stored in the user's web browser (an R-identifier) and stores the user's geographic location on the server along with the ID stored in the cookie. Subsequently, instead of logging page URLs alongside the visiting computer's IP address, the web server can now store the cookie ID and the page's URL whenever the site logs page views.

Apple in this scheme can collect all the information it needs for their stated purposes, which means that from Apple's angle this alternative scheme makes no difference. For the user, however, it does make a difference because cookies are R-identifiers whereas the IP addresses are L-identifiers. Apple therefore in the alternative scheme has no means of locating the named individual that visits its site unless the user allows for a connection by providing her personal data which can be associated with the cookie. Furthermore, if the user destroys the cookie on their machine, this effectively erases any link between user and website, which offers additional protection for the user. These features would, in my view, be clear benefits for the user.[105]

*

8. CONCLUSION

IN THIS ARTICLE I HAVE TOUCHED UPON SOME of the pressing issues of internet use in our times: dataveillance and profiling. Although I have not provided solutions to these issues, I do think that making an explicit distinction between L-, R-, C- and S-type identification may help further in unraveling the complexities of the issues. The main contribution of this paper lies in distinguishing between L- and R-identifiers as identifiers that have different characteristics with respect to their ability to connect to individuals.

With an L-identifier in hand it is possible to go and find the associated individual in the real world. R-identifiers require both the issuer and the individual to present the identifier in order to make the match. An R-identifier therefore in itself is useless to locate the associated individual unless the individual cooperates by means of showing their R-identifier on request or acknowledging a match if the other presents the R-identifier. L-identification serves other needs than R-identification. The goals, relations, issues, and effects differ. Online service providers should consciously consider what kind of identification they require: is a

104. See "Apple Customer Privacy Policy," (2007) <http://www.apple.com/legal/privacy> (emphasis added).
105. Given the enthusiasm with which data retention regulation is being introduced, the prospects for replacing IP logging by cookie logging are not bright. This does not, however, solve all problems. As mentioned, *supra* note 22, IP addresses are also stored in order to prevent "gaming" with the service provider's business model and for security reasons. Whether these aims can be accomplished by means other than logging IP addresses is worth investigating.

connection to named individuals necessary, or does recognition suffice? Depending on this question either an L-identifier or an R-identifier should be issued.

The current regulatory framework, and in my view also the current debate regarding privacy and data protection, conflates L- and R-identifiability into a single concept. This causes confusing debates, puts people on the wrong footing and results in fighting the wrong battles. This is unhelpful in getting privacy advocates and industry aligned. Even though they have different interests in the end, there may be much more common ground than the discussion using the current terminology allows. Separating the various kinds of identifiability and amending definitions and regulations in line with this distinction may help in fighting the correct battles.