Engineering for Privacy
Reducing personal information and complying to privacy law

Michael Sonntag

Privacy is about controlling which data on you is stored by others and what it is used for. There are many different interests about this data: Consumers/Citizens want to reveal no data at all, while companies/government needs it for providing services (or desires more for providing more/better services). Reducing the personal data needed for providing services can ameliorate this conflict. After a short introduction to the EU privacy directive, methods for reducing personal data during gathering/processing as well as transmission are presented. At the end, some thoughts on how to balance the remaining divergent interests are presented.

1.Motivation

There is a large desire for privacy and anonymity intrinsic in every person ([8], [14], [22]). However, this must be limited when the legitimate interests of others are touched. This is therefore a natural field of controversy between different stakeholders [6]. Companies for example want to learn something about their (potential) customers like their credit rating, desires or previously bought products, allowing them better tailored advertising and providing improved service, and through this gaining competitive advantages [2]. However, in E-Commerce there is few or no trust, so customers are unwilling to provide information about them unless it is in their clear interest or unavoidable, regardless of other legitimate interests. Companies therefore sometimes resort to illegal methods for acquiring data (like e. g. Webbugs [7], [17], [24], Cookies [10], [12], [21], Spyware [25], …). Therefore a method for balancing these conflicts is needed. This is usually privacy law, which is standardized within the EU.

This lack of trust is understandable because of past experiences: Everybody often receives uninteresting mail containing private information (like the date of birth or profession), which should not be publicly available. This is also a problem of privacy law, which is quite good, but not anchored in the minds of citizens (awareness is lacking) and only very weakly, if at all, enforced. To receive more trust (and for companies because of this also more information), placing a stronger emphasis on privacy already during design and

implementation is important. This also reduces potential clashes with privacy law. We will take a short look on the EU privacy laws and then discuss how the use of personal data can be reduced during gathering, processing and transmitting data. Emphasis is put on still fulfilling all necessary tasks, so no reduction of functionality takes place.

2.The EU privacy directive

This directive (PD, [4]) applies to all data related to a natural person (in Austria also to legal persons) which is either identified or at least identifiable [13]. This includes any data on computers or software (e. g. agents) related to a single and identifiable person. For certain types of data (race, health, religion, …), more stringent regulations are set. Two important areas can be distinguished: What requirements are defined for the data itself (content) and those for processing this data.

2.1.Requirements for the data

If data is collected, this must be done for a specific, explicit and legitimate purpose. This is important, as changing the purpose later always constitutes a "transmission" (see below). Data must also be adequate, relevant and not excessive in relation to the purposes for which it is collected. This is a criteria of minimalism: Only the data actually needed for the purpose specified may be gathered, and not more: "just in case it might be useful". It must also be accurate (no interpretations and guesswork) and kept this way through continual updates if it is further used. Also, the relation to a certain person may only be retained as long as it is necessary for the purpose, resulting in the obligation to anonymize the data afterwards, if it need not be deleted altogether.

These very strict guidelines possess a large "loophole", however: What the purpose the data is collected/used for is not legally defined and can be set almost arbitrarily by the person or company processing it. A very important step in reducing data is therefore to exactly define what the data is needed for, as only then the elements to be collected and what to do with them over the course of time can be determined. It must also be noted that while the purpose can be set very wide, this also results in large requirements for ensuring the correctness. Also, the organizational problems (ensuring only the persons intended to can access the data for processing it for defined purposes) grow then. For an approach to a technical solution to this last problem see [19].

2.2.Requirements for processing data

If data is gathered or already available and shall be used, this process is also strictly bound to certain requirements. Without fulfilling one of the following, no data may be processed:

- The person it relates to has unambiguously given its consent. Requirements for this are discussed later. This will be the legal reason for processing in most cases.
- Processing is needed for fulfilling a contract the subject is party to; a provision important for E-Commerce (it ends after complete fulfillment; extension is possible for times of warranty).
- Processing is necessary for compliance with a legal obligation of the owner, to protect the vital interests of the data subject, or for the performance of a task carried out in the public interest or in the exercise of official authority.
- Processing is necessary for legitimate interests pursued by the controller or the third parties to whom the data are disclosed (except where they are overridden by fundamental rights of the subject). This must be seen rather strict.

The most prominent consequence from these requirements is, that secretly (without disclosing this to the subject) gathering data is almost never allowed.

2.3.Consent

Obtaining consent is usually required for gathering and processing of personal data. This must be done before any processing starts (this is e. g. a problem with cookies: The homepage may not use cookies, as no information can be given in advance!). The requirements (Art. 2h PD) for consent (which must be given by the person concerned or by an representative) are these (see also [1]; [16] for a discussion what information must be given and on formal requirements):

1. Without compulsion: This may not be seen too narrowly. Providing a service or product only in return for some personal information is perfectly legal.
2. Specific case: Consent can only be given for a specific or a group (which must be exactly defined and understandable for the audience) of uses of data, but not for to generic uses (“We are allowed to do with it what we want” or “will be used for advertising purposes” are invalid clauses).

3. Informed of all circumstances: This especially applies to who will receive the data and where to go to for exercising the privacy-connected rights (inquiries, corrections, deletions, …).

3.Reducing personal data: Gathering

The first step in reducing personal data is avoiding collecting it in the beginning (see also § 3 para 4 TDDSG: principle of avoiding gathering, processing and using data; [11]). The same applies to processing it: Only the data actually needed, and only those of the persons actually affected may be used. Both aspects are determined by the purpose the data will be, respectively is, used for.

If, for example, the URL's the user visits are needed, the straightforward implementation is to just store them. If this is done, a lot of personal and sensitive information, which is unnecessary, is also saved:

- URL's can be of the type ftp and then probably contain the username and the password to access an FTP server. Both are obviously not needed. Are URL's of the type "telnet://" (accessing a remote host; or other protocols) really interesting? Should "mailto:" URL's be stored? This allows checking to which E-Mail address mails were sent, possible including the default subject and/or content! What is to be done with the input to forms, which is also included in the URL if the GET method is used (e. g. "http://www.google.at/search?q=some+searchterms")? What if the user opens two windows? Do we also want to track the sequence of URLs and must therefore detect in which window an URL was requested (or use the referer field)?
- Is it really important to know the date someone requested a catalogue (storing the data; but also providing sequential identification numbers) or his/her date of birth in E-Commerce? Is it necessary to store the IP address of the client in weblogs or isn't a hash-value (allowing the same analysis) enough?
- In E-Government often fees have to be paid. This was previously done with stamps, but now electronic payment is also possible. This results in a lot of additional information received by the administration (or at least the public servants), like which bank you use and your account number there, or credit card information.

The first step for reducing privacy issues is therefore reducing the data collected to the minimum needed for its intended purpose. Special care must be given to by-product data, which is often not gathered but created alongside: Using online questionnaires results not only in the answers to the questions, but also entries

in logfiles, cookies on users computers, or perhaps logs on other servers if e. g. foreign advertisements are included.

3.1.Analysis

The important conclusion from this is, that a detailed analysis is needed which data (or which parts of it) are really necessary, and what other parts could be included, but should better be removed. To help reach this goal, the following checklist can be used:

1. Define an exact specification of the data: This allows removing the unnecessary parts (see below) and helps in parsing it later. The specification should go rather deep and split everything up into the smallest parts possible, even if they are needed for the application only as a whole: There might be subparts in it which are superfluous. Also important is defining the semantics: What is the exact meaning of each element? This simplifies the design of the program and allows deciding on the necessary parts.
2. Data for core functionality: Is the data needed for the core functionality? If yes, we must gather it, otherwise it must be decided whether it can improve our service (in which way exactly, utility for both the customer and/or the company) or not. Only data currently needed should be gathered; data perhaps of use in the future should only be prepared for, but not yet collected.
3. By-product data: Data that is inadvertently created (e. g. serial numbers for customers include partly the date, storing the Austrian social security number includes the date of birth, storing network packets includes the unique numbers of the network interface cards, …). This information should be removed if possible. However, the problem is detecting it in the first place. A step for this is not storing any data, the structure and content of is not exactly known.
4. Personal relation: Must this data be attributed to a certain person or is it sufficient if we know, that this data all belongs to the same person (but we do not know to whom) or some other entity? Perhaps it is sufficient to identify a computer, program, piece of hardware, …?
5. Statistics sufficient: Are statistics enough (collect and sum/manipulate the data and then immediately delete the details) or do we need to store all the individual data? Could storage be avoided by creating multiple statistics (everything possibly needed) at once?
6. Change the default policy: “If we don’t exactly know what it means, just store it” to storing only those parts of the data we exactly know the meaning of and drop everything else. Examples are additional data and options for extensions: Store the options we know about and are interested in, and remove

the rest (this happens in protocols, which very often allow extensions and modifications; probably proprietary ones).

7. Update (+Archiving / Deletion) policy: Keeping the data correct is a legal requirement, though a rather weak one. But it is also in the interest of a company to keep it up to date. What measures will be taken for this, e. g. actually marking addresses where mail was returned as undeliverable, searching for duplicates, or asking customers (to notify the company of changes, allowing them to update it themselves using the web, regular inquiries, …) and what will they cost? Is this worth the utility of the data during its estimated time of usefulness (validity)?

3.2.Methods

There are several methods to improve privacy, even though data is necessary and gathered:

- Pseudonyms/One-way functions: If it is only important to uniquely re-identify a person as the same one, but not who it really is, pseudonyms can be used. This can be a bit tricky, as it is important to distinguish where the association to the pseudonym is stored: If within the company or government, this is in reality the type "partitioning data" (see below). Only if this information is stored externally, e. g. the customers use a pseudonym to log in, this works (see also [9] for pseudonyms in certificates, where the association is stored by a third party with strict rules when to unveil the real identity; or [5] with regards to chipcards). Very similar, but avoiding this problem, are one-way functions. Personal identification is required, but instead of storing this information only e. g. a hash-value of it is stored. Because they are statistically unique, re-identifying the user is possible, but not attributing the data associated with this value to a person. This results in legal easing, as data of this type is only indirect personal data, for which there are several rules less than for fully-identifiable personal data. Pseudonyms possess another advantage if regularly used: The person becomes attached to them and the reputation of the pseudonym gets important the more persons know/recognize it, resulting in more responsible behavior. The drawback is, that the more important this pseudonym gets, the more privacy will be requested for it by the owner.
- Introducing uncertainty: Storing not the complete data or intentionally modifying it to be inaccurate enhances privacy. An example is sorting webpages accessed into groups and storing only them instead of complete URL's. Another solution is modifying data randomly according to statistics. E. g. the mean value of the data stays the same, by virtue of this still allowing statistical analysis. But as nobody knows

which data was modified or to which extent and removing the "noise" is impossible, information on individuals is lost.

- Partitioning data: An improvement, but not a solution, is partitioning the data. There is one part of the data connected to an individual (which is strictly secured with restrictive access), and the majority of the data which cannot be attributed to anybody without the other part. This is only a technical solution (separately storing the association between the person and a random identification number), as legally seen the data is not indirectly personal (which would result in lower protection). The same person or organization still controls both parts and may also combine them if they want to.
- Different entities: It is not always required that a person is identified; often some different entity like an object or a proceeding is sufficient. An example is when paying for some service in public administration: the bank details of the citizen are not needed; only the confirmation that some amount was paid for a certain proceeding. Who did this and in what way is of no concern. So shifting the association from a person to something not related to a person improves privacy.
- Anonymization: If identification of the person is no longer required the data can (and must) be anonymized. This will only be useful as a method itself rarely, as usually the data could be deleted then anyway.

4.Reducing personal data: Transmission

Similar to the previous step, reducing privacy issues can also be done during transmission of data. "Transmission" is understood here as passing the data from the user (or his computer) to the company or creator of the program/application, as well as from the company owning the data to another company wishing to use it once (assuming this is allowed). If the personal information remains on the user's/first companies computer, it is still under his/her control and might yet be useable for the (second) company in some way (e. g. providing personalized services). Also, no (keeping the data under the subjects control) or at least reduced (allowed for use in "unsafe" third countries; if the purpose stays the same, no transmission takes place) problems with respect to privacy laws arise, which otherwise happens if non-EU countries are involved [23].

Some guidelines for design of such aspects are:

1. Encryption of transmission: Personal data should always be encrypted when sent over public channels to fulfill the security requirements. However, basic protocols (e. g. SSL) should be used instead of "homegrown" encryption within the application. This ensures higher quality and also allows the user easier verification of the data actually sent through inserting an intermediate layer before encryption. Also standards (especially open ones and in cryptography) possess their own merits.
2. Keep data locally: Is there a need for retrieving the data from the user's computer? If not, legal problems, archiving and storage is no longer a problem. However, users might accidentally or intentionally delete it. Also it cannot be used for other purposes, even though they were allowed. It is also commercially desirable to possess full control over perhaps business-critical data.
3. Indirect personal data: Not always the exact identity of the person must be known. Might it be sufficient to gather data, but keep only indirect information, or store the information to identify the person on that user's computer? Legally insufficient (but reducing problems) is partitioning the data: A (more or less) public part and a private part (with restrictive access) linking the public parts to a certain person. Transmitting only the former part is less restrictive if done without the latter.
4. Throwing away data: Sometimes it is unavoidable to transmit data (e. g. when a user requests a webpage his IP address, browser version, etc. is sent, even if not wanted). A decision is needed whether this data should be stored (e. g. in a weblog) or be thrown away. This could also be done either immediately or after some processing (e. g. statistics or until the transaction has completed) was done.
5. Using a trusted third party: Is the data needed for continual use or only for backup or documentation? In the second instance, a third party could be used to store the data. This can also be done encrypted, so the archiver has no access to it. Contracts can bind him to only release it with consent of the subject of the data. Legally this is equivalent to deleting the data, but with the option to later recreate it if a dispute or another necessity arises.

4.1. Methods

- Store data in cookies: One method is storing the data on the users computer in cookies. This is dangerous for its existence as they can be deleted at any time without notice. Also, their size is restricted and rather small (4 kByte), allowing only few data to be stored.
- Separate application on users computer: Storing the data on the users computer instead of at servers has several advantages, like the data remaining under control of the user, no legal problems, and that

the same instance can be used on different servers, easing maintenance by the user. This would work well in combination with existing technologies like P3P [18]: The user stores data and sets a privacy policy. If the server matches it, the data is automatically sent or used (consent is provided through the P3P configuration by the user [15]). However, there are some disadvantages also: Companies do not have any access until contacted by the user, which is unsuitable for many applications (e. g. shipping goods) and problems arise if several computers are used (synchronization!). This is also no guarantee for privacy: If the user can be recognized as the same visitor as before, a profile of his activities can still be created. Also, P3P does not help enforcing the policy [3], it is just a declaration.

- Passing "usage" only instead of data itself: From the result mostly the same, but legally different is, whether the personal data itself or only its usage is sold to another company for a certain use. In the former case the new owner receives full control over the data for a certain purpose and can use it within this boundary for their intentions, but also assuming full responsibility for it (correctness, inquiries, etc.). In the latter case, the control and the responsibilities remain with the original owner, restricting the spread. This does not require the owner of the data to do all the work: The data can still be passed on to the other company, but it must be deleted after its use. This is therefore mostly a legal differentiation and appropriate for "one-shot" use, like sending advertisements.
- Use pseudonyms: The data on the person is stored locally, but no identification of the person is possible (resulting in indirect personal data). The data stored must be exactly defined to not contain any unique references (e. g. E-Mail address). See also above.
- Used third party for storage: The whole personal data is stored by a third party, which is trusted by both sides. The disadvantage of this approach is, that the company cannot use any of the data in any way, even for providing services like e. g. personalization in the interest of the user. It is therefore only appropriate if the data is required only as a safeguard (evidence).

5.Balancing privacy issues

For increasing the trust of customers an explicit privacy policy is needed. It is important that this should neither be a legal nor a technical document, but using the "language" of the customers so they can understand it. This includes explaining the methods (e. g. what is a "cookie"?) used for collecting the data, what the data will be used for, and to whom it might be passed on (all legal requirements, but they can be fulfilled

in many ways!). And this policy must then be strictly enforced. This is a very important part, as changing it also causes a lot of problems: Data must be handled according to the policy in effect at the time it was created or gathered, requiring labeling it. As often is used on websites "We reserve the right to change our privacy policy without notice; the customer must go look for changes; continued use constitutes consent" might only be applicable for future data, but certainly not for previously collected data. Consent requires knowledge of all circumstances and applies only to a specific case: Future unpredictable changes of the policy are no specific case, so no consent could have been given when the policy was accepted.

Another important factor is education of employees on privacy. Explicit information what may be done with information is needed. This also allows them to better inform customers on inquiries. E. g. quickly looking up some data on persons privately known to them needs to be seen as amoral instead of as a minor and forgivably gentlemen's crime.

Some education is also needed on the part of the consumer or citizen. If a company has proven to be trustworthy (past experience, seals, policy, …), information should be revealed a bit more freely and more accurately. Often questionnaires are filled in deliberately wrong, which is also a source of problems (bad service, inappropriate offers, etc.) for them, but attributed to companies instead of themselves. Especially in the web the customer will have to acknowledge that there is no "free lunch" anymore. If valuable and personalized services are consumed, some payment must be made: if not in money perhaps by information about them (which, in itself, is not bad, only the use made of this information is commonly the source of the problem). Often this information might be valuable to companies (and in the end also for the customer) while there is no appreciable loss because of revealing it. The main danger materializes when different companies correlate and integrate their data. Therefore partitioning the data and enforcing strict rules on transmission are very important (see also the problem of data transfer form the EU to the USA; Safe Harbor Agreement). This also addresses the needs of companies to acquire data on their customers. If they receive a bit more of it freely, there is no or at least very much reduced need for obtaining it illegally or from other sources (possibly gaining more data then, than actually needed). Therefore in my opinion privacy should shift a bit from protecting persons from revealing their data, to avoiding sharing and interrelating data, as only the second is the really dangerous threat. This is, at least partially, also the intention of the German TDDSG with it partitioning rule and prohibition of bringing together data from different services [20].

6.Conclusions

When comparing privacy to security, some parallels can be drawn: Permissions usually are granted on the base of “need to have access for fulfilling the work”. Information should be provided based on “need to know to provide (better) services”. Both are in most danger when connected to another domain, which is under different control. Therefore some ideas can probably be transplanted: Concentrate on security to the outside, as the inside is under different control (contracts, employees are more trustworthy than outsiders); this is closely related to enforcing rules for transmission (including change of purpose) of data in contrast to gathering it. Another example is changing the security policy: This is an important task and extensively evaluated for even remote consequences; changed privacy policies can also lead to repercussions in different areas and should not be modified just on a whim.

We must not forget that privacy is only a means and not an end to itself. Both of the extreme positions often encountered have their arguments, but a compromise is needed. This compromise should consist of providing more information to companies, but more/better restrictions about sharing this information with other institutions.

In the short run however (and perhaps generally), methods for reducing the use of personal data should be employed: an exact analysis of the data actually needed and how collecting as well as passing it to others can be avoided.

7.References

[1] BÜLLESBACH, A.: Datenschutz be Data Warehouses und Data Mining. CR 1/2000, 11

[2] BUXEL, H.: Die sieben Kernprobleme des Online-Profiling aus Nutzerperspektive. DuD 25 (2001) 10, 582

[3] CAVOUKIAN. A., GURSKI, M., MULLIGAN, D., SCHWARTZ, A.: P3P und Datenschutz. Ein Update für die Datenschutzgemeinde. DuD 24 (2000) 8, 475

[4] Directive 95/46/EC of the European Parliament and of the Council 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. Official Journal of the European Communities L281, 23.11.995, 31. http://www.datenschutz-berlin.de/gesetze/europa/den.htm (2.6.02)

[5] ERMER, D. J.: Systemdatenschutz und Chipkarte. CR 2/2000, 126

[6] GRIMM, R., ROßNAGL, A.: Datenschutz für das Internet in den USA. DuD 24 (2000) 8, 446

[7] HAAS, F. G., SOSNA, A.: Was versteht man unter WebBugs? http://members.surfeu.at/privacy/definitions/webbugs.html (2.6.02)

[8] HILLENBRAD-BECK, R., GREß, S.: Datengewinnung im Internet. DuD 25 (2001) 7, 390f

[9] HOPP, C., GRÜNVOGEL, A.: Pseudonyme nach dem deutschen und österreichischem Signaturgesetz. Datenshcutzrechtliche Aspekte rechtsvergleichend betrachtet. DuD 26 (2002) 2, 79

[10] IHDE, R.: Cookies–Datenschutz als Rahmenbedingung der Internetökonomie, CR 7/2000, 417

[11] IMHOF, R.: One-to-One-Marketing im Internet – Das TDDSG als Marketinghindernis. CR 2/2000, 110

[12] JAHNEL, D.: Datenschutz im Internet – am Beispiel des Speicherns von Cookies. In: SCHWEIGHOFER, MENZEL, KREUZBAUER: Auf dem Weg zur ePerson: aktuelle Fragestellungen der Rechtsinformatik. Wien: Verlag Österreich 2001

[13] JAHNEL, D.: Datenschutz im Internet. ecolex 2001, 85

[14] KÖHNTOPP, M., KÖHNTOPP, K.: Datenspuren im Internet. CR 4/2000, 253

[15] LOHSE, C., JANETZKO, D.: Technische und juristische Regulationsmodelle des Datenschutzes am Beispiel von P3P. CR 1/2001, 55

[16] OLG Frankfurt: Haushaltsumfrage zu Werbezwecken. Urteil vom 13.12.2000 – 13 U 204/98 (LG Darmstadt), CR 5/2001, 294 (with comments by NILS, L.)

[17] OPSAHL, K., INFANTINO, S.: Privacy and the Use of Pixel Tags. http://www.perkinscoie.com/resource/ecomm/pixel.htm (2.6.02)

[18] P3P: http://www.w3.org/P3P/ (2.6.02)

[19] PETERS, F., KERSTEN, H.: Technisches Organisationsrecht im Datenschutz – Bedarf und Möglichkeiten. CR 9/2001, 576

[20] RASMUSSEN, H.: Datenschutz im Internet. Gesetzgeberische Maßnahmen zur Verhinderung der Erstellung ungewollter Nutzerprofile im Web – Zur Neufassung des TDDSG. CR 1/2002, 36

[21] SCHAAR, P.: Cookies: Unterricthung und Einwilligung des Nutzers über die Verwendung. DuD 24 (2000) 5, 276

[22] SCHAAR, P.: Persönlichkeitsprofile im Internet. DuD 25 (2001) 7, 383

[23] SIMITIS, S.: Der Transfer von Daten in Drittländer – ein Streit ohne Ende? CR 7/2000, 472

[24] SMITH, R. M.: Web Bug FAQ. http://www.eff.org/Privacy/Marketing/web_bug.html (2.6.02)

[25] VAMOSI, R.: What is spyware? http://www.zdnet.com/products/stories/reviews/0,4161,2612053,00.html (2.6.02)