Minimizing the Risks of Data Protection Infringement
Data Lifecycle Risk Assessment

Silvia Balaban[1] and Manuela Wagner[2]
[1]*FZI Forschungszentrum Informatik, Haid-und-Neu-Str. 10-14, Karlsruhe, Germany*
[2]*Center for Applied Legal Studies, Karlsruhe Institute of Technology, Vincenz-Prießnitz-Str. 3, Karlsruhe, Germany*

Keywords: Legal Challenges of Big Data in Multi-discipline Services, Privacy Impact Assessment, Anonymization, Data-lifecycle-evaluation.

Abstract: In this paper, we propose an evaluation scheme which has the objective to permit the user to identify the legal data protection obligations through a continuous data-lifecycle-assessment-method and to re-design the data processing actively. To ensure the compliance with data protection principles under the European law and thus preventing the risk of sanctions, it is necessary, especially in multi-discipline services, to continuously check during the complete data-usage-process whether personal data are given and which methods of risk minimisation like the application of anonymization techniques are useful.

1 INTRODUCTION

Innovative research concerning predictive maintenance or prediction of population behaviour is expected to contribute to major efficiency-improvements especially in the areas of transport and energy. Big Data analytics in multi-discipline services increase the expectation of an improved networking and cross-linking of information deriving from different sources. Those innovations might conflict with fundamental principles of data protection law. As the EU General Data Protection Regulation (GDPR) will be applicable from May 25th 2018, controllers (persons processing personal data for own purposes) need to adapt their data processes in order to comply with new GDPR-concepts and avoid tighter sanctions. This emphasizes the need for a practical guide for controllers without legal background. As the GDPR is not applicable in case of anonymous data, it might be useful for controllers to continuously check whether personal data are still required. However, controllers must be aware of the risks that datasets become personal as capabilities identifying natural persons develop. We therefore propose a risk-assessment-method, which permits controllers to cope with the obligations of the GDPR and simultaneously provides them the evidence of fulfilling the duties. As legal obligations are not negotiable, data protection risk management relates to a balancing of risks and safeguards rather than costs (Friedewald et al, 2016). This paper shows the opportunities of a risk management, based on existing work on Privacy Impact Assessment (PIA). To cope with typical Big Data challenges, we propose risk minimization, pointing out hindrances and benefits of anonymization. To estimate the necessity of safeguards we conclude with a practicable Data Lifecycle Risk Assessment.

2 RELATED WORK

Firstly, we present risk-assessment-methods and focus on the privacy challenges of Big Data.

2.1 Privacy Impact Assessment

The GDPR obliges controllers to carry out a PIA prior to the processing, only in cases when the processing likely results in a high risk to the rights and freedoms of the data subjects. PIA is an instrument to recognize and evaluate the potential impacts and risks caused by data processing and is a possibility to reveal unintentional data protection risks (Friedewald et al, 2016). The Britain Information Commissioner's Office (ICO) released a generic, technology-neutral PIA-model (ICO, 2014), while the Commission Nationale de l'Informatique et des

Balaban, S. and Wagner, M.
Minimizing the Risks of Data Protection Infringement - Data Lifecycle Risk Assessment.
DOI: 10.5220/0006358103560362
In *Proceedings of the 2nd International Conference on Internet of Things, Big Data and Security (IoTBDS 2017)*, pages 356-362
ISBN: 978-989-758-245-5

Libertés (CNIL) provided recommendations and good practices (CNIL, 2015). The GDPR requires for a PIA a processing description, an assessment of necessity and proportionality in relation to envisaged purposes, related risks to data subjects' rights and freedoms and countermeasures including safeguards and security measures (Art. 35 para 7 GDPR). Additionally, the Standard Data Protection Model, which extends the information security principles CIA with data protection goals, provides evaluation schemes but seems to assume that personal data are already given (SDM, 2015). As this will be one of the key questions for controllers, a simplified testing scheme distinguishing personal and anonymous data would be useful. Our paper focuses on how to prevent privacy risks in the "legal grey area" between personal and anonymous data in context of Big Data.

2.2 Big Data Privacy Challenges

Although there is no fixed term for Big Data, the legal debate is based on general assumptions like the big scale of data collection including high variety and detail of the collected data and the aim to combine data from many different sources (Art. 29 WP, 2013). Based on this scenario several conflicts with current and future data protection law have been already pointed out. At this point, we describe the typically assumed infringements with basic data protection fundamentals.

2.2.1 Legitimation

Processing of personal data requires a legitimation, which could be a legitimation by law or the data subjects consent. The big scale of data collection might conflict with the requirement of an individual case-by-case-assessment, when balancing legitimate interests between controller and data subject (Ulmer, 2013). A consent is considered as valid, if it is freely given, specific, informed and unambiguous. Legal experts claim that the capacity of discernment is missing in regard of data processing in multi-discipline domains and that declarations of consent are often too indefinite (Kamp and Rost, 2013). Without equivalent alternatives data subjects might feel forced to give their consent in order to get a relevant service (Brummund, 2014).

2.2.2 Purpose Limitation

Legitimate data processing must be based on specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes. The greater the risks, the more specific purposes must be outlined beforehand (Art. 29 WP, 2013). This obligation is challenging if a combination of data from different sources is envisaged or if an exploratory analysis is conducted to develop the research question (Bornemann, 2015). If Big Data analysis is built on (cross-border) linking of personal data from various contexts, the impacts and inferences drawn from the data are hardly predictable (Raabe and Wagner, 2013). "Non-linkability", a principle presented in the Standard Data Protection Model (SDM, 2015), clearly contradicts the idea of re-using and connecting existing data through different disciplines for new purposes (Weichert, 2013).

2.2.3 Storage Limitation, Data Minimisation

Personal data may only be processed, if necessary and proportionate for attaining the specific legitimate purpose (Bizer, 2007). This is the case, when there is no less severe possibility, which is adequate to reach the focused purpose. An assessment must be made in each respective case to comply with privacy by design and by default. It is assumed that disproportionality is given, in the case of collecting and storing as much data as possible to provide an extensive basis for data mining also for potential future purposes (Roßnagel, 2013).

2.2.4 Fairness, Transparency, Accuracy

Only with sufficient information data subjects can understand and control decisions they are subject to and decide autonomously about exercising their rights (Art. 29 WP, 2013). Parties must be aware of the processing purpose or context before the collection (Ohrtmann and Schwiering, 2014). In principle, the right of information, correction and erasure or restriction must be granted. Enforcing these rights depends on determining the responsible controller(s) and means to contact him/her. Concerning Big Data in multi-discipline services with multilateral processing chains the variety of controllers and processors lead to hardly bearable challenges especially concerning the possibility to effectively enforce rights (Raabe and Wagner, 2015). The combination of information deriving from different sources might also lead to unintentional findings. Personal data must be kept up to date and incorrect data should be erased or rectified without delay. Concerns were raised that decisions based on data mining might be inaccurate or discriminatory, as the algorithm draws conclusions based on detected correlations. These statistical inferences may perpetuate existing prejudices and stereotypes (Art. 29 WP, 2013).

2.2.5 Integrity and Confidentiality

Personal data must be processed by ensuring appropriate security. Expanding data volume might lead to a rise of security threads (Weichert, 2013).

3 BENEFITS AND HINDRANCES OF ANONYMIZATION

The presented potential conflicts between Big Data scenarios and data protection principles does not mean that Big Data is illegal in general. Besides the fact, that a case-by-case legal assessment is always necessary, there are possibilities to limit the contradictions. One possibility to be excluded from the scope of data protection law is to anonymize, although the legal dispute how to distinguish personal and anonymous data must be respected. Using anonymization techniques is still beneficial to comply with the GDPR by mitigating certain obligations.

3.1 Distinguishing Personal and Non-personal Data

To ensure whether a planned data processing requires a legitimation the controller should analyse if personal data are given. The qualification when data is considered as personal is highly controversial and was subject to a recent decision by the European Court of Justice (ECJ, 19. October 2016, C-582/14). As the upcoming GDPR provides the utmost similar wording like the current Data Protection Directive, it is expected that the courts key considerations will remain relevant (Kühling and Klar, 2017). The testing scheme shown in figure 1 helps the controller to identify if his data are of personal nature. If the dataset contains information about a natural person, it must be verified, whether this person is directly or indirectly identifiable. If the data is primarily related to objects (or events), it must be analysed whether "multi-relations" to an object and a person are given: depending on the proximity, information related to the object might also provide information about the person, who is e.g. using or owning the object (Art. 29 WP, 2007). A whole dataset might be "infected" by reference information (Weichert, 2007). The same situation occurs, when object-related-data is collected with the aim to link them to a person (Forgó and Krügel, 2010). Following this it has to be checked whether the person is directly or indirectly identifiable which means either the controller or a third party is in the possession of means of identification without disproportionate effort of time, cost or manpower. In this regard, information of third parties has only to be considered, if accessible for

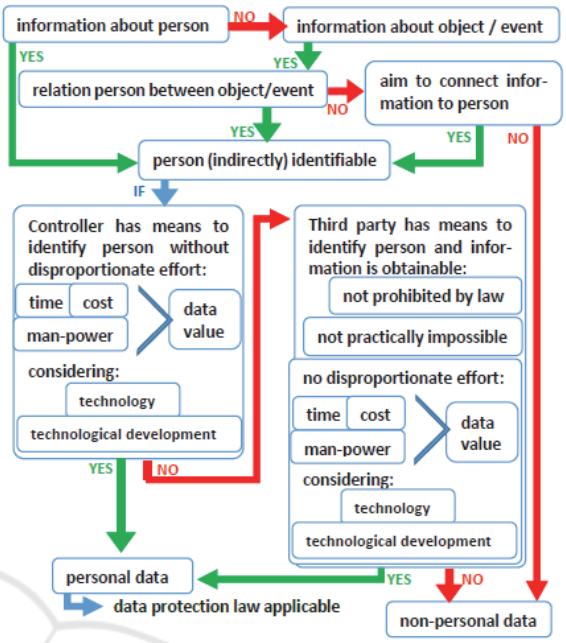

Figure 1: Testing scheme for personal data.

the controller, which is not the case, if this is illegal or impracticable. The question of estimating the proportionality of identification effort is still subject to legal discussion. Considering the court decision it is unlikely that the individual value of identification for the specific controller can be taken as differentiation criterion as the ECJ bases the qualification of personal data on relative and objective criteria (Kring and Marosi, 2016). The term relative comprises the possibilities and accessible means of the individual controller. To achieve a certain level of certainty, an objective determination of the efforts proportionality is preferable. Otherwise, the qualification of personal data and applicability of a legal framework depends on financial resources or economic priorities of individual controllers. It can be assumed that most data comprise a certain value. Disproportionality is given, if the effort exceeds any potential value of identification and an attempt of identification cannot be expected under rational circumstances (Weinhold, 2016). The risk of achieving additional identifying information by third parties is considered insignificant, if impracticable because of disproportionate effort or illegal. It is subject to discussions whether data access possibilities can be assumed in case of the theoretical existence of a legitimating norm (Kring and Marosi, 2016; Jensen and Knoke, 2016; Weinhold, 2016;

Kühling and Klar, 2017) or if the preconditions of this legitimating exception must be fulfilled in the specific case (Eckhardt, 2016; Kartheuser and Gilsdorf, 2016). The better arguments plead for the first opinion, as it must be predictable whether data is personal or not. In the current ECJ-case, a cyberattack or an infringement of copyrights would have enabled the request to submit additional data. Then random encounters will influence the applicability of data protection law and already stored data might change to personal data. (Kühling and Klar, 2017)

3.2 Technical, Organizational and Contractual Measures

Anonymization might be a good procedure to avoid data protection infringements. Due to technical innovations in the field of Big Data, anonymization is hardly realizable as the risk of re-identification by combining and analysing data from different sources rises. Big Data for predictive maintenance focuses on machine data, but still challenges occur in keeping non-personal data anonymous. Terms like k-Anonymity, l-Diversity or t-closeness focus on the determination whether a dataset is anonymized but without considering potential combinations with additional information. By combining various databases, it was possible to de-anonymize a k-anonymized dataset (Narayanan and Shmatikov, 2008). When the data is still considered personal, the processing is subject to data protection law, imposing sanctions to controllers lacking the necessary legitimation for the processing. Using further technical, organisational and contractual measures (TOC) can impede the re-identification effort leading to disproportionality. Normally, technical and organizational measures are undertaken to strengthen security, but could also be focused on the exclusion of identification. Examples for measures ensuring anonymity are given in Table 1.

Table 1: Examples of measures to create or ensure disproportionate effort of de-anonymization.

Technical measures	- encryption - data usage control - adding noise - availability, access control - functional separation of data - data obfuscation
Organizational measures	- documentation - training - audits
Contractual measures	- confidentiality / non-disclosure agreements - notification obligations

A first step is the selection of relevant data sources, identification of obtainable additional information and evaluation which safeguards can prevent identification. Contractual obligations alone can never be considered as obstacle to identification as parties can renounce the contract (Dammann, 2014). However, in combination with technical and organizational measures identification risks could be minimized. This could guarantee the non-applicability of data protection law. If the data are still considered personal, TOCs could have a positive effect as risk minimizing safeguards to fulfill the data protection obligations. The anonymization can be considered as "partial" from a legal point of view.

3.3 Risk Minimisation

In the cases where anonymization is only partial this still has an impact of the legitimacy of Big Data scenarios as it can work as a safeguard compensating high risk data processing.

3.3.1 Legitimation

Art. 6 para 1 (f) GDPR can legitimate data processing through balancing conflicting interests. Safeguards like partial anonymization can mitigate potential negative impacts of the processing and therefore influence the balancing significantly.

3.3.2 Purpose Limitation

The GDPR provides the possibility to change purposes if a compatibility assessment has been accomplished, concerning different aspects like reasonable expectations of data subjects or potential impact. The tests outcome of the test also depends on the „existence of appropriate safeguards, which may include encryption or pseudonymisation" (Art. 6 para. 4 (e) GDPR). Safeguards can also be technical and organisational measures, partial or full anonymization or aggregation of data to prevent any inappropriate impact on the data subjects. This factor may compensate a change of purposes, which otherwise would fail the compatibility assessment, or compensate for a lack of clear specification of purposes. (Art. 29 WP, 2013)

3.3.3 Storage Limitation, Data Minimisation

If partial anonymization limits the level of information retrievable by processing personal data, controllers could comply more easily with the requirements of necessity and proportionality. Art. 25 GDPR provides the concept of privacy by design/

default and mentions pseudonymisation and data minimisation as potential measures. A partial anonymization is also a way of data minimisation, as less personal data are processed.

3.3.4 Fairness, Transparency, Accuracy

Art. 11 GDPR foresees a case where a controller is not able to identify the data subject but meanwhile personal data is given as the GDPR is applicable. Meeting requirements of information would create irresolvable problems for controllers, therefore the data subjects' rights of access, rectification, erasure, restriction and data portability do not apply.

3.3.5 Integrity and Confidentiality

Art. 32 para. 1 (a) GDPR states pseudonymisation and encryption as a risk minimizing possibility. Partial anonymization might have the same effect like pseudonymisation.

4 DATA LIFECYCLE RISK ASSESSMENT

These findings show the relevance of designing privacy risks based on an overall risk assessment. As every change of the envisaged processing operations may have an impact on the outcome of the legal evaluation, a frequently effectuated check of the legal challenges concerning Big Data scenarios should be foreseen. The PIA could be used as a reference model. The first question arising is whether personal data are involved. So, we propose the combination of the PIA with a first evaluation concerning personal data and, if necessary, a legal evaluation of all data protection principles. One relevant case for constant re-assessment is the case of potential personal data where the processing would comprise a high risk in the moment an identification of natural persons is possible. A one-time assessment only prior to collection would lead to the conclusion that in absence of personal data no data protection obligations occur. Most important to notice is the fact, that the qualification of information as personal data might change over time, as the obtainable knowledge or the accessible technologies improve. In this case, an anonymous database might be considered as containing personal data, even if no further data is added to this database. A frequent analysis of a data-lifecycle from collection to deletion is therefore mandatory. An overall approach for Big Data in multidiscipline sectors should combine a continuous risk assessment concerning risks of identifying natural persons with the risk assessment concerning the impact on these persons' rights and freedoms if the probability of identification is given.

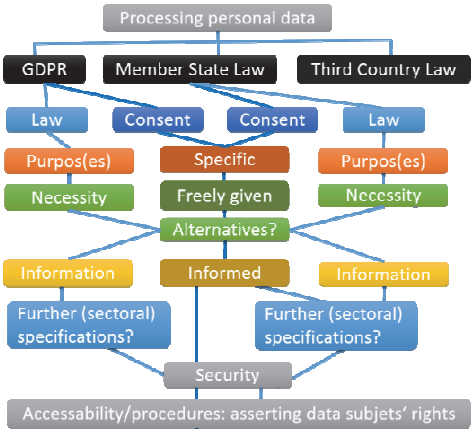

Figure 2: Generic overview of data protection obligations.

As already lined out in Section 2 the obligations from data protection law merely stem from the data protection fundamentals. To give an overview of the legal challenges under the GDPR the first legal evaluation might be based on data protection key factors shown in figure 2 to assess the effort of compliance or usefulness of anonymization. A re-design of the envisaged lifecycle could be either aiming on the avoidance of personal data or the fulfillment of these general protection principles. Controllers should use a systematic assessment order to document and trace the undergone analysis following the steps in figure 3:

Data-Lifecycle-Scenario-Identification: specify the relevant scenario from the data collection to the deletion and list involved parties.

Obtainable Information enables identification of natural persons (data subjects)? An evaluation is necessary if datasets contain personal data focussed on the outlined lifecycle scenario and must be repeated frequently, especially with every new processing step or new obtainable information.

Legal Overview: a first evaluation based on the data protection principles shown in figure 2 provides an overview of generic requirements and shows the probability of infringements.

Lifecycle-Re-Design: to anonymize/prevent the emergence of personal data or minimise risk to comply legal requirements

Anonymization ensured? If a data processing should not fall under data protection law a frequent test of the probability of re-identification is recommended to avoid infringements.

Review legal Analysis: if personal data is given, a full legal analysis must be done, additionally a PIA is necessary in case of a high risk for the data subjects' rights and freedoms.

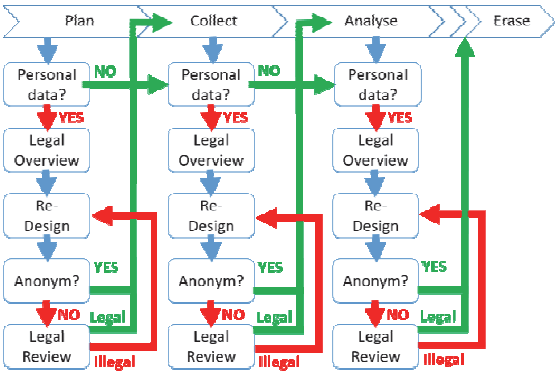

Figure 3: Approach of a data lifecycle risk assessment.

5 CONCLUSION

For Big Data scenarios conflicting with fundamental data protection principles anonymization is one solution, but the re-identification risks have to be taken into account. Nevertheless also partial anonymization can be considered as safeguard compensating data protection risks and therefore might legitimate big data scenarios in multi-discipline services. By a lifecycle risk assessment controllers are able to estimate whether their data is anonymous or which risk minimizing efforts are necessary. Future work should be focused on technical implementation. A tool, which could automatize the testing scheme and thus permit the controller to check whether anonymization is still given and show him corresponding duties and respective safeguards would be therefore highly valuable for future research.

REFERENCES

Art. 29 WP, 2007. *Opinion 4/2007 on the concept of personal data.* Available at: http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2007/wp136_en.pdf. Consulted on April 10 2017.

Art. 29 WP, 2013. *Opinion 03/2013 on purpose limitation.* Available at: http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf. Consulted on April 10 2017.

Bizer, J., 2007. *Sieben Goldene Regeln des Datenschutzes,* Datenschutz und Datensicherheit, pp 350-356.

Bornemann, D., 2013. *Big Data – Chancen und rechtliche Hürden*, Recht der Datenverarbeitung, pp. 232-235.

Brummund, A., 2014. *Smartphones und Apps: Datenschutzrechtliche Risiken und deren Begrenzung*. In *GI-Jahrestagung*. pp. 539–550.

CNL 2015. Privacy Impact Assessment. Available at: https://www.cnil.fr/sites/default/files/typo/document/CNIL-PIA-2-Tools.pdf. Consulted on April 10 2017.

Dammann, U., 2014. In *Simitis Bundesdatenschutzgesetz*, Nomos, Frankfurt , 8th edition.

Eckhardt, J., 2016. *Anwendungsbereich des Datenschutzrechts – Geklärt durch den EuGH?*, Computer und Recht, pp. 786-790.

Friedewald, M. et Al., 2016. *Datenschutz-Folgenabschätzung*, White Paper, Forum Privatheit. Eggenstein, 2nd edition.

ICO, 2014. Conducting privacy impact assessments code of practice. Version: 1.0. Available at: https://ico.org.uk/media/for-organisations/documents/1595/pia-code-of-practice.pdf. Consulted on April 10 2017.

Forgó, N., Krügel, T., 2010. *Der Personenbezug von Geodaten – Cui bono, wenn alles bestimmbar ist?,* Multi Media und Recht, pp 17-23.

Kamp, M., Rost, M., 2013. *Kritik an der Einwilligung*. Datenschutz und Datensicherheit, pp 80–84.

Kartheuser, I., Gilsdorf, F., *EuGH: Dynamische IP-Adressen können personenbezogene Daten sein,* Multi Media und Recht, 382533.

Knoke, F., Jensen, S., 2016. *EuGH-Urteil zur Personenbezogenheit dynamischer IP-Adressen: Quo vadis, deutsches Datenschutzrecht?,* ZD-Aktuell, 05415.

Kring, M., Marosi, J., 2016. *Ein Elefant im Porzellanladen – Der EuGH zu Personenbezug und berechtigtem Interesse*. Kunst und Recht, pp 773-776.

Kühling, J., Klar, M., 2017. *EuGH: Speicherung von IP-Adressen beim Besuch einer Internetseite, Anm.,* Zeitschrift für Datenschutz, pp 24-29.

Narayanan, A., Shmatikov, V. (2008). *Robust de-anonymization of large sparse datasets*. In Security and Privacy, 2008. IEEE Symposium, pp. 111-125.

Ohrtmann, J.-P., Schwiering, S., 2014. *Big Data und Datenschutz – Rechtliche Herausforderungen und Lösungsansätze,* Neue Juristische Wochenschrift, pp 2984-2989.

Raabe, O., Wagner, M., 2015. *Verantwortlicher Einsatz von Big Data*, Datenschutz und Datensicherheit, pp. pp 434–439.

Roßnagel, A., 2013. *Big Data – Small Privacy?,* Zeitschrift für Datenschutz, pp 562-567.

SDM, 2015. *Das Standard-Datenschutzmodell*, V.0.9. Available at: https://www.baden-wuerttemberg.datenschutz.de/wp-content/uploads/2015/10/SDM-Handbuch_V09a.pdf. Consulted on April 10 2017.

Ulmer, C.-D., 2013. Big Data – Neue Geschäftsmodelle, neue Verantwortlichkeiten?. Recht der Datenverarbeitung, pp. 227-232.

Weichert, T. 2007. *Der Personenbezug von Geodaten,* Datenschutz und Datensicherheit, pp 113-119.

Weichert, T. 2013. *Big Data und Datenschutz,* Zeitschrift für Datenschutz, pp 251-259.
Weinhold, R., 2016. *EuGH: Dynamische IP Adresse ist personenbezogenes Datum – Folgen der Entscheidung für die Rechtsanwendung,* ZD-Aktuell, 05366.