EJLT European Journal of Law and Technology

Self-made data protection – is it enough? Prevention and after-care of identity theft

by Oliver Vettermann[1]

Abstract

The Cambridge Analytica scandal was not the first data breach that shattered users' expectations of internet services. Many online services had similar incidents before, based on flaws in IT-security or caused by a member of the company. All of them have one thing in common: the data breach leads to a massive risk for personal data and the user's identity. Further, the consequent damage could hardly be limited or stopped because data can be copied and shared infinitely. Therefore, preventive services for identity leaks has been created – so-called identity leak checkers. These services certainly help to protect individuals' digital identities. However, the legal and technical aspects are rarely discussed. This article analyses these leak checking services and outlines their positive and negative aspects. Then, focussing on the negative aspects, it outlines guidelines for a revised version of an identity leak checker. Including both legal and technical aspects, this revised version would lead to "Effective Information after an Identity Theft" (EIDI) – also the name of an actual German research project with the same acronym explained in detail in the article.

Keywords: data protection, data breach, digital identity, identity theft, prevention, leak checker.

1. Introduction: The terminology of data breaches, leaks and digital identities

Cambridge Analytica, Uber[2], Yahoo[3] and LinkedIn[4] have something in common: they are victims of data breaches and have had to defend their vulnerabilities to external audiences. In other words, their technical infrastructure – either software or hardware – had a flaw in IT-security which was used to steal (personal) data. As a consequence, control over the "original" data was lost – as outlined in Art. 4(12) GDPR.

Every data breach reduces the trust of internet users in Big Data companies as well as the security on other websites. In turn, the risk of consequent identity theft as well as consequent financial or reputation damage rises. This latter effect in particular not only eventuates because of a data breach, it occurs as a result of the subsequent leak of stolen data. After a data breach has happened, the loss of control results in the disclosure of the stolen data by hackers. Data breach collections or combined lists (so-called Combo-Lists) are then shared publicly via PasteBin or Twitter – or similar services – and some of them are even shared in hacker boards. No matter where the data is published, it can be used to steal an individual's digital identity and impact their rights and freedoms.

The following article will concentrate on leaked digital identities. To explain: a digital identity generally consists of personal or personalised data and may be considered as similar to a personal profile. But the definition is not focused on the personalisation of data. Rather, it requires an aggregation of data much more similar to a profile. So, an anonymous digital identity might also be covered by the definition even if only personal data is covered by data protection law, see Art. 1(1) and Recital 26 GDPR. The difference between anonymous profiles and pseudonyms, however, is in their use: some digital identities are necessary to present websites in a convenient way and to deliver (digital) services more fluently and purposively. Regularly, these identities are "filled" in by the user when he/she registers on the site for the first time. In these cases, a digital identity has the purpose of identification, so the device or user can be recognised across multiple browsing sessions – this enables, for example, content to be consistently displayed to subscribers. If this use is based on a user profile, the digital identity corresponds with the definition of a personal profile when personal data is linked (or linkable for the service provider) to the user. This is usually the case when an online service includes a payment model. On the contrary, some digital identities are created "in secret" by giving a device a certain ID aimed at allowing identification of a user and at displaying the most relevant – at least according to the algorithm used – personalised advertisement. In this way, it is also possible to track individuals between different browsing sessions by tracking their online behaviour – their so-called "browser fingerprint". Prominent examples for these kinds of IDs are browser cookies and IP addresses. They can be connected to profiles and log-files located at the backend of a website. These tend only to be used for authentication because it is irrelevant who you are (for identification) – it is only relevant what you need, watch or click on. Whether the digital identity is secret or not, both forms still constitute digital identities: both consist of aggregated data and, by combining or collecting data, a piece of your (digital) identity is represented as data. This definition has its origin in the more technical terms of identity management systems, where the focus is on collection and linking personal data.

To round out the explanation: the highest risk after a data breach persists for leaked credentials of a digital identity, which includes personal data like your home address or banking account. With this data, digital identities are the aim of phishing attacks and data breaches, because the credentials constitute valuable goods.

2. Overview: After-breach provisions

In order to prevent and fight against the above discussed undesirable scenario, internet/website users can take measures to protect their data. Tools or actions such as email encryption or two-factor-authentication of the accounts, however, can only prevent certain breach possibilities and depend on the support of the website being used. For example, not every website offers a two-factor-authentication.[5] But these features are merely part of the technical and organisational measures outlined in Art. 25 and 32 GDPR and are not described in detail in this article. In particular, there are tools for the phase between the data breach and the leak of data. The objective of these tools is simple: the data subject is informed of the likelihood of being affected by a breach and what he/she should do afterwards to prevent a possible misuse after a leak. The following websites or projects – also called identity leak checkers – have been designed with the intention of implementing parts of this paradigm.

a) Have I Been Pwned and similar tools

The best-known international representative for identity leak checkers is the website "Have I Been

Pwned".[6] Immediately on the homepage, one finds the entry prompt for the email address you want to check. After entering a valid email address, the leak database of the service is scanned for the entry and gives you a warning red or calming green return. Especially, if your email is found in the database, the website tells you where your leak data was from. However, there is no detailed information about legal aspects or what to do next as the subject of a data breach. Since March 2018, the website simply recommends you to use 1Password (a password manager) to better protect your passwords.[7] In addition to that scanning mode, you can also check for leaked passwords or for "pwned" (which means "hacked" in hacker language) websites. The return message follows the same warning-calming-scheme, as well as general information concerning the protection of credentials.

A similar leak checker to "Have I Been Pwned" is "BreachAlarm".[8] Other international/English identity leak checkers are "Has my email been hacked?",[9] "Leaksource"[10] and "Leakbase".[11] These work in exactly the same way as "Have I Been Pwned". They are not, however, recommended as a result of doubts concerning their data sources.

b) Germany: HPI Leak Checker/ BSI Security Test

The language is not the only difference here: the HPI Leak Checker[12] and the BSI Security Test[13] work in a different way to the leak checkers discussed in the previous paragraph. Accordingly, it is necessary to explain them a little more detail.

The HPI Leak Checker is a service to check whether your email address appears in the leak database of the Hasso-Plattner-Institute. Compared to Have I Been Pwned and other comparable services, there is an email prompt here too. The difference, however, lies in the more detailed results of the leak database scan: while the positive response confirming no leaks still constitutes a short positive message, the negative response confirming the presence of a leak shows you your level of infection much more clearly. A chart, divided into a variety of data categories, shows which category is infected and the possible match between the leaked data and the initial data breach. Following this information, the service gives you some recommendations about what you can do if your password or bank account information has been leaked. Psychological support – indeed any kind of psychological element – is still missing in the message. Accordingly, the recipient of the email could still panic after reading breach results. Also, you can only check for leaked emails, but not for passwords or other kinds of data. Further information concerning the data leak sources is missing, too.

Less detailed is the BSI Security Test, which is a leak checking service from the German Federal Office for Information Security (BSI). After several investigations by law enforcement authorities regarding bot networks and identity theft, sixteen million digital identities were found to have been compromised.[14] Thus, federal authorities set up this leak checking service to scan these data sets and to warn and protect German citizens. This action is based on major data breaches and numerous identity thefts investigated by the public prosecutor. To steal the identities, many emails and passwords were stolen and used for contracting under a different name as well as for other illegal uses. In terms of the identity leak checker itself: again, the prompt offers a possibility to check your data against the tool's database. The response from the tool is given via email, but is not as detailed as the HPI leak checker. Only the abstract wording in the FAQ for the BSI leak checker website tells you that there will be recommendations for after-care and for the prevention of further damages. These recommendations seem similar to the twelve general advices on how to protect your digital devices against attacks on the internet.[15] A more detailed screenshot, or explanation, is missing. Like the other leak checking tools, you get an email as response and warning – this means you only get a mail when your data was found to have been breached.

Otherwise, there will be no response, which seems quite difficult from a psychological point of view. For example, it could nonetheless be necessary to hand over certain recommendations to protect your digital identity from theft or other misuse. Not to forget, this lack of information also pertains to the leak source – which data breaches are included in the database – or further information about the response. The only fact you can find is that every user of certain email address providers might be infected, e.g. t-online, gmx.de and Vodafone.

c) Switzerland: MELANI leak checker and SwissLeak

Similarly to the German system, Switzerland also has a Reporting and Analysis Centre for Information Assurance (MELANI).[16] MELANI's task is simple: protect natural persons – namely data subjects as described in Art. 4(1) GDPR – but also small companies from data breaches through warnings and advice. In general, the authority offers a selection of information about current attacks across the internet. Furthermore, it is possible for users to register phishing mails and websites[17] or report other kinds of misuse[18]. To optimally use the analysed data, MELANI also offers a check tool comparable to the leak checking services discussed above.[19] If an email address or – novel in comparison to other checking tools – username is entered, the checker offers results instantly on the website. Additional information about a positive result and what to do when your data is compromised seems not to be provided and there is no information in the frequently asked questions section. In this case, the general recommendations in the section "How do I protect myself?", especially the rules of conduct, can and should be obeyed. A psychological element is missing here too.

A completely different "experience" is the service provided by SwissLeak.[20] SwissLeak has scanned millions of leaked datasets for data from Swiss citizens and redacted them into a more understandable and visual form. For example, you can zoom in and out on a map of Switzerland to see the companies from which leaked data has come as well as where data breaches themselves happened. In particular, both companies of public interest as well as federal authorities are included. Thus, the database of SwissLeak contains a broad picture of leaked data. The tool provides a detailed view concerning the data breach situation in Switzerland as well as risks and technical vulnerabilities. A deeper look reveals critical personal data and possible high risks in data protection law. In fact, SwissLeak collects **full** anonymised credentials. But during the analysis of them, the passwords are unhashed (or decrypted) at a certain state, which leads to a high risk of misuse and possible attacks on SwissLeak itself. As the website describes freely for every entry, there is a counter for any decrypted password – which literally means plain-text passwords. As a consequence, SwissLeak has access to the digital identities of data subjects and further, data is personalised and identifiable in the most cases. Like the research section of the SwissLeak website shows, 362,577 inhabitants of Zurich are affected by data breaches and 28,361 digital identities are personalised enough to be identifiable in relation to specific individuals. If SwissLeak suffers a data breach and the unhashing process is misused by hackers, the risk of damage is enormous. European data protection law in this case may be not helpful in every case, because Switzerland is not part of the European Union.[21]

d) Conclusion

The above mentioned identity leak checking services/tools establish the possibility of self-made data protection and help with the after-care – or at least with the recognition – of the consternation of an identity theft. In general, this provides some degree of protection. But to access this protection requires a certain knowledge of the handling of personal data and credentials, how to identify a trustworthy identity leak checker website as well as the use of the tools regularly – if you can't get full protection by using a password manager like 1Password.

Regular users or owners of digital identities are unlikely to be at this level of knowledge – maybe because of the hurdles in understanding the technical aspects of digital identities, identity theft as well as how personal data can be misused. In this regard, the awareness of the value of data is not growing much in relation to the internet. An indicator for that is the fact that the annual ranking of the worst passwords still refers to the simplest and/or funniest choice: "hello" or "123456".[22] Leak checking services are doubtless important, but only digital natives and so-called "nerds" are using them.

To reach the average user of digital identities as well as well-informed users, it is necessary to evolve and expand the model of current leak checkers and build up a leak checking service version 2.0. This conclusion not only relates to the missing knowledge of average users. Other important facts which have to be considered in order to develop a better leak-check model are the lack of IT-security in the software framework and the missing psychological aspect. For instance, SwissLeak saves the full credentials, the email addresses and usernames as well as the passwords, in hashed strings. But the service is able to decrypt them and to verify how many usable accounts their leak database consists of. In contrast with European data protection law, this is not a legitimate way of using/ storing data: Similarly to the GDPR, the Swiss data protection law includes on data minimisation and purpose limitation principles, see Art. 4 Paragraph 3 DSG (abbreviation for the Swiss data protection law). However, it does not have effective penalties as outlined in Art. 84 GDPR. Further, it sounds doubtful to process found leak data and encrypt these without knowledge of the affected data subjects. For further information about the shown leak profile of a company, there is still the possibility of a right to access the data, though this is only possible in exchange for a fee. Considering the purpose of data protection law itself, this payment model totally violates the aim of data protection and leads to a right for data protection only for the wealthy. Again, this is not what European data protection law – with regard to Art. 15 GDPR – should aim to achieve, even if Switzerland is not one of the Member States of the EU. This approach of an identity leak tool seems to fail completely.

Regarding other tools, the effort involved in preventative and self-made data protection is another problem in the leak checking service approach. Most of the tools don't offer an information service if your data is found in a consequently later leaked database. That's why an individual with multiple digital identities has to test every email address or password regularly - and by hand. Also, this has to be done on every website mentioned, because there is no clear information about which leak database is included in each service's database. This challenging task could never be done without a kind of app or significant knowledge on the part of the individual.

In summary, an identity leak checker version 2.0 should go in hand with a high IT-security level and a Zero-Knowledge-Protocol to protect the data subject as well as the processing company, regardless whether it is a controller or a processor.

3. The effectiveness of after-care concerning identity theft and general recommendations

At first sight, the owner of a digital identity can protect his/her data against attacks and misuse by using a variety of different tools. To comply with the GDPR, the leak checker should also follow other aims of EU data protection law to achieve and maintain the intended level of data protection.

Leak checker companies are also bound to the obligations and conditions of the GDPR if they process personal data. This is problematic, because not all data from an IT-security breach is

personal data as the term is defined under Art. 4(1) GDPR. To give an example, leaked data could contain email addresses and passwords as well as credit card numbers – but this differs in every leaked database. If a leak connects a credit card number with a name, the number is personal data in terms of Art. 4(1) GDPR. Otherwise, it is an identifiable pseudonym which is personal data too, depending on the likelihood of identification. For the identity leak checker as controller or processor under the GDPR, every data from a leak is therefore personal data. As Recital 26 elaborates, it seems reasonable and possible that a leak checking service could identify natural persons because of the connection to several public leaks. However, the likelihood of identification also could depend on the data itself and corresponding rights to receive further information.[23]

With regard to the principles of data protection in Art. 5 GDPR, there is a lot more to consider when regarding the processing of personal data during leak-checking. In general, the data subject has to be informed concerning the use of his/her data. This obligation is concretely outlined in Art. 14 GDPR. If data leaks are found and processed by a leak checker to build up a leak database, the data subject should be informed, if Art. 14(1) GDPR and the principle of purpose limitation and transparency are to be followed strictly. But informing data subjects is not always possible. Usually, publicly leaked data is barely part of the area of responsibility of the data controller. So, the data subject does not know who is processing the data – the original data controller or the controller who's responsible for the website with the leaked data – and how it is processed exactly. In these cases, Art. 14(5)(b) GDPR contains exceptions for situations when informing a data subjects leads to a disproportionate effort or is impossible. In this case, or if the data is not personal data in terms of Art. 4(1) GDPR, there is no obligation to inform the data subject. As shown, the recovered data is personal data in general because of the reasonable possibility of linking it with a single individual. Even if the effort is unreasonable in light of the mass of recovered leaked data and the large number of affected data subjects, the GDPR still mandates that information shall be provided by public notice as an appropriate measure to protect the data subject's rights and freedoms and legitimate interests.[24] Also, if the data subject can be related to the credentials at a later time, for instance by combining further leak databases, the data subject must be informed without considering the exceptions of Art. 14(5) GDPR.

Additionally, during the processing and aggregation of leak databases, data not necessary for information and possible investigation of the data subject should be deleted during the first collection and processing of the data. This requirement is related to the principle of purpose limitation and data minimisation in Arts. 4(1)(b) and (c) GDPR respectively.

Touching on the processing of leak checking services again, a legal basis/ foundation is required to process data lawfully. Such bases are exhaustively outlined in Art. 6 GDPR, from which Art. 6 (1)(a), (b), (c) and (e) or (f) could be helpful to find a lawful interest for leak checking services.

Article 6(1)(b) GDPR justifies processing regarding the performance of a contract. For example, a data subject can contract with a processor to protect credentials or other data like popular password managers do. Sometimes, this service is intertwined with an insurance against (digital) identity theft. In the German insurance market, for example, many well-known insurances offer this option in their portfolio. These options have in common that data transferred by the data subject is compared with publicly available data. An analysis of their general contract clauses shows their lack of data protection and integrity, because most of the data is handed over to a processor or other third parties who do the whole protection and analysis part in the contract. Even if it is explained in the contract – somewhat surreptitiously – the collection of personal data by a third party and out of reach for the insurance company doesn't seem transparent or

trustworthy if this information isn't told directly to the insured in their portfolio. What is also not explained clearly are the clauses for changes in purpose as well as the missing individual opt-out possibilities. Both are imposed by the principles of purpose limitation and data minimisation in Art. 5(1)(b) and (c) respectively. Further, to reduce the trust in these kinds of insurances, several insurances are related to the same third party responsible for fulfilling the insurance service. This leads to a major risk of an identity theft by hacking into identity databases that already exists as well as an inter-connected database with profile data for each insured person. But, if there is a contract to which the data subject has agreed freely, a lawful reason for processing is possible in general if the contract is based on leak checking. This rationale corresponds with Art. 6(1)(a) GDPR as well as the freedom of contract. Nonetheless, the rules of the GDPR regarding processing personal data must still be taken into account, especially Art. 24 GDPR.

In contrary to the above, Art. 6(1)(c) and (e) GDPR both lead to a special legal obligation for the data controller. Art. 6(1)(c) GDPR outlines a legitimation of processing following from a national law and therefore corresponds to a direct obligation on a processor. Similar to this, Art. 6(1)(e) GDPR sees lawful processing may happen if processing is in the public interest and/or in the exercise of official authority vested in the controller. In both cases, a processing with a legal basis in European or national law is required according to Art. 6(3) GDPR. The obligation in national law as mentioned in Art. 6(1)(c) GDPR relates to the more private law aspect, whilst Art. 6(1)(e) GDPR is focused on a task concerning the public interest. As well as the German national data protection law – the BDSG – the Regulation doesn't contain an explicit obligation considering the rights and interests of a company. Regarding the obligations to the purpose of public interest, the UK Data Protection Act, for example, provides rules relating to official authorities and public interests.[25] These rules also explain that the term "public interest" isn't meant to be a blanket clause which could be filled with any common interests a society might have – e.g. IT-security or interoperability of different messaging systems.[26] Accordingly, leak checking as a precaution in IT-security is not included even if it is necessary for user data in the digital age. As well as the concrete obligation in law, the public interest has to be interpreted as part of governmental function. So, Art. 6(1)(c) GDPR requires the fulfilment of tasks by the government through authorised private companies. To conclude, leak checking could be a public interest if the aim is the protection of digital identities of all natural persons (Recital 14 S. 1 GDPR). Therefore, private companies could also be authorised to help protect and check databases. However, data protection law has to be followed as explained in the EIDI software framework. Otherwise, the intended protection will in fact turn into a major risk for personal data.

The most discussed legal processing ground is that outlined in Art. 6(1)(f) GDPR, which allows processing because of a lawful interest pursued by the controller or processor. This wording could be interpreted in a very broad way, so the processor could understand it to include every substantial interest.[27] To limit this interpretation as an exception, the interests of the company shall not override the rights and freedoms of the data subject in the balancing test. In fact, the interests of the company have to override or be in balance with the data subject's interests.[28] The interests of the data subject can be found in the fundamental rights and freedoms such as the right to the protection of personal data (Art. 7) and the right to respect for private and family life (Art. 8) in the Charter of Fundamental Rights of the European Union (CFREU). Further references concerning legitimate interests can be found in Recitals 47 to 49 GDPR. On the other hand, a legitimate interest of a company is based on (concrete) lawful, economic or non-material reasons covered by Art. 15 and 16 CFREU. Regarding data protection, companies have a legitimate interest in protecting their network and information security by processing personal data if it is necessary and proportionate for the purposes of ensuring IT-security – see Recital 49. If the processing is

"necessary to guarantee the security and continued proper functioning of the online media services that it makes accessible to the public" a legitimate interest can be found in saving and processing IP-addresses to protect the digital infrastructure against DDOS (Distributed Denial of Service) attacks.[29] To lead back to identity leak checking, the interest of IT-security depends on the detailed purpose of the service. When the leak checking service is used to get more information about data subjects through their digital identities, this is against the principles of data protection – e.g. lawfulness, fairness and purpose limitation. In addition to that, illegitimate uses of data are accompanied by consequences in criminal law. Equally, processing data during identity leak checking for IT-security reasons is affordable but bound to the principles of data minimization and purpose limitation. This is permitted by the wording in Art. 6(1)(f) GDPR, when the "processing is necessary" and "proportionate for ensuring [...] security".[30] As explained, further data from the data subject which is not needed to analyse whether an existing digital identity is affected has to be deleted and must no longer be processed. Under these circumstances, identity leak checking is a possible tool for providing a high level of IT-security. More detailed limitations and conditions depend on the processed data and the leak checking service, especially the algorithm.

It should also be mentioned, but will only be briefly discussed here, that Art. 6(2) GDPR may constitute another lawful reason to process personal data. This, however, only refers to national regulation and data protection law in relation to Arts. 6(1)(c) and (e) GDPR. Next to – already discussed – Art. 6 of the (UK) Data Protection Act, further exemptions corresponding to intelligence services' processing (e.g. automated decision-making) can be found in Art. 111 and Schedule 11 of the Data Protection Act. Concerning further investigation, these lawful interests are less important than the explained ones.

In conclusion, leak checking is possible under the GDPR but bound by certain limitations. Every processing of personal data has to comply with the GDPR – starting with the principles in Art. 5(1) and the lawful interests in Art. 6(1), (2) GDPR. Moreover, standards in IT-security must be followed to protect leak checkers' digital infrastructure and to prevent checkers themselves from becoming aggregated database of digital identities. Lastly, this risk has to be avoided in general.

4. The research project EIDI: A game changer?

So far, leak checking has not been serviced in a detailed and extensive way. Services lack in providing assistance with protecting individuals' data in future as well as regarding which tools should be chosen concerning legal and technical aspects. The agitated data subject is left on his/her own. With a glance at current European data protection law, one sees that a data protection by design and by default approach as a way of identifying leaked identity data is necessary.

To this effect, the by the Federal Ministry of Education and Research has chosen to support the Effective Information after a Digital Identity Theft – EIDI – research project. This project intends to find a better solution for both data subjects and affected companies.

4.1 The scope of the project

The research project is focused on building an effective leak checking service that provides a warning function by checking leaked data against actual data sets next to an advisory function in the software framework. In this regard, different project partners are observing and researching legal aspects (FIZ Karlsruhe), data protection aspects of software frameworks (Independent Centre

for Privacy Protection Schleswig-Holstein, DPA), psychological aspects of warning and after-care in identity theft (University of Duisburg, Research Team for general psychology and cognition) and blending these aspects into source code for a bespoke software framework (University of Bonn, Fraunhofer Institute). To follow data protection law and other legal limitations, only publicly available data can be checked and processed by the framework. The term "public" has different meanings regarding leaked data, because of the different kinds of published sources which are usable: Fully publicly available data can be found without hurdles by searching on free available databases like PasteBin. On this website available datasets contain every type of data, not only credentials or credit card numbers – sometimes also a full copy of identity data. Similar to that, there is partly-publicly available data, which means available after overcoming minor obstacles. Typical examples of this are internet boards where published data is only available after login but without payment. Apart from these boards, there are also black-market boards where digital identities – or sets of identities – are traded. This kind is explicitly not in the scope of the project, to avoid infringement of both criminal and data protection law.

Further, the data should be communicated between the leak checking service holder and the software framework which is set up by the leak checking company itself. For example, a company which administrates a social network could implement the software framework in the infrastructure, whilst the framework itself updates and checks in the background – maybe supervised by IT-security experts and data protection officers. The checking data then would be distributed by or with the help of a holder as a central authority. Because the leak checking service collects and sends the data for checking, the algorithm should anonymise or pseudonymise the credentials to ensure and implement a Zero-Knowledge-Protocol and should neither store nor log unencrypted credentials to reduce the risk of damaging the whole framework infrastructure.

4.2 Technical issues

Technical standards have to be state-of-the-art and implement data protection by design and default (Art. 25 GDPR). In general, the risks for rights and freedoms of natural persons should be avoided at any time software is being programmed. Typical implementations of data protection by design are the anonymisation and pseudonymisation of personal data, a data minimisation approach during processing and storing data, storage limitation, transparency regarding processing and limited access to personal data.[31] In case of the EIDI software framework, certain key aspects will be discussed here.

As mentioned above, the software framework itself is planned to be used by local (digital) identity providers or companies storing identity data – like banks or online shops. For this use, it is required to hand over the found data and offer a check with the company database for affected identities. Accordingly, the checking mechanism should be based on a Zero-Knowledge-Protocol to protect personal data during this data exchange. The main purpose of Zero-Knowledge-Protocols is that the communication between sender and recipient doesn't contain any secret information. Simply put: the purpose could be compared to the game "Riddle riddle ree, what do you see?". [32] For example, the publicly available password shouldn't be transmitted clearly and passwords in general should be hashed or encrypted for IT-security reasons. As well as this, the answer from the company must be encrypted and should only contain statistics – e.g. how many users have been affected, but not which users. Otherwise a hacker could eavesdrop the communication and use the information for hacking purposes such as using transferred digital identities to login to individuals' accounts and change their passwords in seconds. Then identity theft would still be possible. Therefore, a message in a Zero-Knowledge-Protocol never contains a password or other

parts of the credential but, for example, a randomised number which is automatically included in the password sent by the user.

There are some steps prior to the data exchange between the applicant company and the central authority of the database: the data has to be gained from the leak sources and aggregated in some way for further processing. Unnecessary data has to be deleted in advance because of the principle of data minimisation in Art. 5(1)(c) GDPR. Following this principle, there should be no aggregated datasets of digital identities. Otherwise, this huge amount of digital identities will lead to a high risk of hacking the EIDI-infrastructure as well as potentially risking all connected devices when a man-in-the-middle-attack is used. This would risk trust in the software framework and in the research project. However, these scenarios are not desirable and thus, the source of the checking databases should be organised in a decentralised manner or by the applicant company itself. For example, the already checked databases could be logged to avoid mistakes or endless warnings. The logfile itself then only has to be only readable by the software framework – similar to an asymmetric cryptography system (e.g. public-key cryptography). Further, it shouldn't contain any of the checked data itself but maybe only hashed sources of leaking sources combined with a duplicate filter. Because of this, maybe the central authority of in the EIDI software framework communication should be avoided and the framework should work on its own.

Ergo, there are some ideas concerning the implementation of appropriate technical measures that should find their way into the code of the software framework. Until the final framework is ready, actual measures are always considered during the supervision of framework construction – especially regarding the application of companies in their role of being a data controller and/or processor in data protection law.

4.3 Legal issues

Mainly, there is a gap to bridge regarding the legal issues when it comes to identity leak checking. This interdisciplinary question has to be solved on multiple terrains, although the main role seems to be taken by data protection law.

4.3.1 Constitutional law and fundamental rights

First and foremost, the German constitution is, in certain aspects, the superior instrument in the legal hierarchy regarding formal law. Whenever it comes to the application of law by a public authority, fundamental rights have to be considered and balanced when they collide with public interests. Regarding European law, the CFREU as well as the European Convention on Human Rights (ECHR) have to be included in this balancing process to ensure national law concurs with international law. This modus operandi is ruled subliminally in Art. 23(1)(1) and (2) of the German constitution (GG), which rules on competence regarding the transfer of rights from the state to the European Union.

To get to this state, fundamental rights must have been engaged in the first place. In this regard, fundamental rights from the German constitution which might be engaged are the general personality right in Art. 2(1) combined with Art. 1(1) GG – especially the right of informational autonomy and the right to guarantee the confidentiality and integrity of informational systems – and telephone/ data exchange secrecy in Art. 10(1) GG or the right to privacy (which is limited to homes or apartments as place for retreat) in Art. 13(1) GG. These rights cover the path of data from their origins to their (temporary) storage. In terms of digital identity and leak checking, the identity is thus protected from a user's first login or registration to the exchange when the identity

is used or has to be transferred to a data controller/ processor.

This kind of coverage doesn't lead to full protection, but to the task for the legislator to take data protection into account – not only concerning digital identities – in formal statutes. The concrete task of legislation arises when a risk to injure a fundamental right emerges between private subjects (e.g. data controller and data subject), in contrast to the relation regarding public authorities or the state. In their case, fundamental rights have to be considered constantly because of their binding effect regarding public authorities and every other institution of the state. To force the state to regulate the current situation between private subjects to overcome actual flaws by legislation, the above mentioned risk has to come up concretely and the possibility for a collision of private goods has to be quite alarming or nearly colliding. Especially if the collision already happened or still remains, a claim against the state could be made before the Federal Constitutional Court – although this is extremely rare. Following this, and to finalise discussion of this aspect of law, just some of the collision situations should be outlined in more detail.

A typical situation where regulation of colliding freedoms by law is necessary is the one between the data controller and the data subject. Because of its imbalance, especially if the controller is an information service or social network – which tend to be monopolies –, the rights and freedoms of users as natural persons should be given greater weight than the company's interests. This is not to say, of course, that every interest of a company is less important. In case of data protection, building up security measures or processing personal data necessary for security or performing the service is also important. Besides that, the legislator has to draw the line between personal and professional data very precisely, which also influences the relevant balancing test. Further, rights and freedoms have to be protected in relation to each other as well – for example between users and other users or between users and hackers – every one of them is a natural person, whose behaviour is not directly regulated by the German constitution. These connections have to be regulated in a very general way by law and more precisely in a range of different kinds of law. For example, to protect a user from identity fraud or data theft, the legislator regulated the use of personal data via data protection law as well as outlining punishments for infringements via criminal law. Furthermore, private law regulates the extent of rights to private autonomy regarding the consent and use of data by companies. Next to this, public law defines voluntary standards for IT-security in Germany, which are then concretised in catalogues of measures provided by the Federal Office for Information Security.[33]

In summary, the (German) constitution – as well as the CFREU and the ECHR – build up a framework of interests and rights which have to be considered in legislation. The details of these laws will be explained in own sections.

4.3.2 Data protection law and the GDPR

With reference to the connection between the data subject and the controller (and maybe also a processor), there is no way to address software framework lawfully without following European data protection law. But, as explained above, the processed data has to be distinguished between personal and non-personal data from the outset. Whereas a pair of ID and password seems to be anonymous, the personal aspect of an email address is undeniable. The first variant has to be considered in detail, because not every pair of credentials is anonymous: As long as the identity hoster/ service provider just generates an ID without processing any personal data – such as full name, address, birth date etc. – the ID or full credentials tend to remain anonymous. On the contrary, if personal data is processed in combination with the ID or username, the ID works as pseudonym between the user and the provider (which can be a data controller or processor) and

is only anonymous for other users on the platform. This relative approach to the personalisation of data leads to two major assumptions: 1) Every data can be personalised data, if combined with other datasets. 2) Full anonymity is rarely given and is nearly impossible to achieve.[34]

Regarding the leaked identity data in the research project, nearly all data contains a full name (usual for personal mail addresses) or is connected to other personal data like full addresses or online-banking accounts. Also, a password can contain personal data, if names and birthdates are built into them – e.g. Emily03-12-1997 contains obviously a date and possibly a birthdate from Emily, the user's daughter. This complexity seems vague but takes into account the variety in possible data usages as well as the problems of big data and aggregation. Finally, this opens up the scope of application of (international) data protection law.

Applying data protection law is the most important aspect of law regarding the software framework. Many rules and guidelines have to be considered during the construction of the software and may be implemented afterwards if a data protection impact assessment should be conducted by the applying company, because the EIDI framework is still processing personal data in the matching process. In general, the principles of data protection in Art. 5 GDPR should be considered throughout the source code. Several specific rules must be considered in processing and in the use of publicly available data, because data protection itself doesn't depend on the geographical location where the data is available or stored. Data protection persists, if the data is stolen or published without consent of the data subject. This thought is also implied in Arts. 33 and 34 GDPR, when the control over the personal data is lost and the data is misused. Otherwise, the rights of the data subject would expire afterwards, which be in contradiction with the purpose of data protection principles such as the right to erasure in Art. 17 GDPR. Besides that, the GDPR and national data protection law – like the German BDSG and the changes by the IT-security Act and Council Directive (EU) 2016/1148 – define a standard in IT-security and elaborate technical criteria in data protection law. The obligations in Art. 25, 32 GDPR ensure a technical state of the art and help to build soft- and hardware with data protection by design/default to comply with the principles of "data minimisation", "purpose limitation" and "storage limitation" of Art. 5(1)(b), (c) and (e) GDPR respectively. Also, the rights of the data subject force technology companies to consider their software regarding the right to erasure and rectify data (e.g. the erasure and rectification problems in blockchain) as well as the possibilities to guarantee transparency and free consent. Lastly, international and national data protection law, as specific law, implement general fundamental rights and interests into technical requirements and thus into the processing software framework. Thus, fundamental rights are widely considered and are referred to in the case of balancing interests like in Art. 6(1)(f) GDPR.

Therefore, and to close the chapter on data protection: data protection outlines the requirements that the software framework has to follow from the first to last lines of code. From the beginning, the project is supervised by legal researchers and experts from a national data protection authority. All the discussed aspects are steady parts of interdisciplinary discussions and lead to a strong implementation of encryption and IT-security measures. Further, the concept of after-care with psychological support via text or by phone (if possible) has also to be legally supervised. In short: Programming software and data protection are constant companions to each other.

4.3.3 Private law and copyright law

To regulate the legal positions of private individuals in German law, the legislator uses private law. Usual situations ruled by private law include contracts or the law of property. Regarding the digital environment, private law serves the base for contracts concerning the use of data as well as

relevant terms and conditions. In general, private law regulates basic claims between private individuals – which means natural persons as well as companies. One of these claims is a claim for omission and elimination to undo any damage done to legal protected rights, especially fundamental rights. So, the general clause of Arts. 1004(1) and 823(1) of the German Civil Code (BGB) opens up on the discussion of the above-mentioned fundamental rights: if any data is stolen, the "owner" of the data (in data protection terms: the data subject) can make a claim against the thief to delete every available leaked data (as far as he/she can) and to refrain from further theft future. In addition to that, the data subject can mount a claim against the breached company if the data breach is based on a significant lack of security as well as further risking the rights and freedoms of natural persons – similar to the Art. 82 and 1(2) GDPR.

Further, copyright law is part of private law and rules on the special conditions of copyright licenses, the limited use of works and on which works are protected by copyright law. Referring to the digital environment, German copyright law contains special rules for databases since the implementation of the European Database Directive 96/9/EC into national law. Databases are protected – only – as a collection of independent works, data or other materials arranged in a systematic and methodical way. Further, they are individually accessible by electronic or other means and should not be confused with computer programs.[35] As the definition describes, the difference between the intellectual creation of the work: whereas the intellectual output regarding a computer program is reflected in the source code and logical connected design patterns (e.g. in computer games), the performance of a database as a work in legal terms is represented in the kind of aggregation and/or arrangement of the data. So, the data itself is not protected by the database definition in copyright law.[36] Instead, if the data is source code or text with an intellectual essence, separate protection – such as that provided to literature or text – could be possible.

Regarding the EIDI software framework, the focus of the investigation is on credentials as independent data. Following the definitions above, the question is: are credentials also protected by copyright law? A brief look at the definition in Art. 1(3) and 2 in Council Directive 96/9/EC gives a negative answer. This excludes independent data *expressis verbis*. Further, the German legislation is implementing this argumentation into national law like in Art. 3(1)(d), 3A Copyright, Design and Patents Act as well as jurisprudence which follows this perspective. As a consequence, only the complete digital identity could be included under the definition when the storage constitutes a database with some unique arrangement or similar criteria.

Besides that, leaked credentials are relevant for claims concerning omission and rectification of damage. As explained in general, the data subject could make different claims against the hacker, the identity seller or the company which publishes the leaked datasets containing the data subject's credentials. Further, a claim against the company where the leaked data comes from seems possible, but it could be overruled by specific or superior law like data protection law, e.g. Art. 82 GDPR.

So far, the EIDI software framework should consider the claims of deleting credentials from the publicly available internet and thereby prevent data subjects from being hacked. Whereas the latter is also a task in data protection law, the former could be arranged by using a connection to the API of a service provider, where possible, to report illegal content. However, the question of illegality here is subject to uncertainty should be considered in further research.

4.3.4 Criminal law

Lastly, German criminal law is also engaged in the building process of the software framework.

Even if the purpose of criminal law does not correspond with IT-security law, it supports and punishes the damage emerging from the insufficiency of IT-security measures. Certain offences in German criminal law punish the unauthorised access to computer material or the digital communication between computers. Also, the data integrity is covered by German criminal law and its rules are similar to those concerning criminal damage. Both German laws are comparable to the rules outlined in the UK Computer Misuse Act. In addition to these rules, the range of punishment is extended by the fines in Art. 83 GDPR. All together, they require IT security measures or "borders" like firewalls and other – literally – data protection measures. Unsecured data, e.g. without encryption or a firewall, is not covered by any legal protection.

Because of these requirements, it is necessary to implement IT-security measures in the source code of the EIDI infrastructure where the data is processed. But these measures are already required to be implemented because of the rules in Art. 25, 32 GDPR and further requirements under data protection law. To summarise, connection between criminal law and the project is rarely explicit but should not be underrated: for example, as a side question, the legal aspects of Honeypots are barely discussed, it could be useful to follow the data streams through the internet and to understand how and where leaked datasets are shared.

5. Summary

Identity leak checkers are not new to all internet users, but are new to most of them. Even if they seem useful for self-made data protection, there are some risks connected to their use. Despite the fact that the well-known international identity checking platform "Have I Been Pwned" seems secure, it doesn't offer help besides the recommendation to use 1Password. In contrary to these quite safe solutions, the MELANI service seems to represent the complete opposite when leaked data is collected, decrypted and available in plain text, at least for a certain time. Even if the decryption process could be helpful for checking data for duplicates, or for understanding hashing algorithms, the decryption of data itself seems to be an infringement of the rights and freedoms of affected persons and companies.

Having these flaws in mind, leak checking services should evolve and implement better data protection techniques including those already ruled by the wording in Arts. 25 and 32 GDPR: "Taking account into the state of the art [...]". For example, stored data should be hashed – and never decrypted in any way – as well as not combined, aggregated or otherwise linked with other personal data in line with the principle of data minimisation. Further, every data transmission has to be encrypted and should be based on a Zero-Knowledge-Protocol.

To achieve and evolve this standard, the project EIDI was founded and is funded by the Federal Ministry of Education and Research. As mentioned, the variety of research subtopics and questions is huge, but in law many of these questions are not without an answer. Even if just some of the key facts could be explained, the interdisciplinarity of the project and the benefits of the EIDI software framework show which aspects of identity leak checking have to be – and can be – improved with the technical, psychological and legal state of the art. If the project is successful enough, an European successor could follow.

References

Books

Ahlberg, H and Götting, H-P (April 20, 2018), BeckOK Urheberrecht, legal commentary (Munich: C.H.Beck).

Albrecht, J P and Jotzo, F (2017), Das neue Datenschutzrecht der EU (Baden-Baden: Nomos).

Kühling, J and Buchner, B (2017), DS-GVO and BDSG, legal commentary (Munich: C.H.Beck).

Journal Articles

Robrahn, R and Bremert, B (2018), 'Interessenskonflikte im Datenschutzrecht: Rechtfertigung der Verarbeitung personenbezogener Daten über eine Abwägung nach Art. 6 Abs. 1 lit. f DS-GVO', ZD 7/2018.

Rossnagel, A and Scholz, P (2000), 'Datenschutz durch Anonymität und Pseudonymität - Rechtsfolgen der Verwendung anonymer und pseudonymer Daten', MMR 12.

Case Law

Patrick Breyer v Bundesrepublik Deutschland [2016] ECJ C-582/14.

Volkszählung, German Federal Constitutional Court [1983] 1 BvR 209/83 – also known as BVerfGE 65, 1.

Statutes

Council Directive 91/250/EEC (Computer Program Directive).

Data Protection Act 2018.

General Data Protection Regulation (EU) 2016/679.

Links

1Password (February 22, 2018), 'Finding Pwned Passwords with 1Password'. Retrieved from: https://blog.agilebits.com/2018/02/22/finding-pwned-passwords-with-1password/ – last accessed October 15 2018.

BSI (January 21, 2014), 'Millionenfacher Identitätsdiebstahl: BSI bietet Sicherheitstest für E-Mail-Adressen'. Retrieved from: https://www.bsi.bund.de/DE/Presse/Pressemitteilungen/Presse2014/Mailtest_21012014.html – last accessed October 15 2018.

BSI, 'Zwölf Maßnahmen zur Absicherung gegen Angriffe aus dem Internet'. Retrieved from: https://www.bsi-fuer-buerger.de/BSIFB/DE/Service/Checklisten/Massnahmen_gegen_Internetangriffe.html – last accessed October 15 2018.

Grossman, L (December 19, 2017), ,'The Worst 25 Passwords of 2017'. Retrieved at: http://time.com/5071176/worst-passwords-2017 – last accessed October 15 2018.

Lang, L (December 22, 2016), ''hallo' ist meistgenutztes deutsches Passwort – auf Platz zehn steht 'ficken''. Retrieved from: https://www.heise.de/security/meldung/hallo-ist-meistgenutztes-deutsches-Passwort-auf-Platz-zehn-steht-ficken-3579567.html – last accessed October 15 2018.

Press Association (May 18, 2016), 'Hacker advertises details of 117 million LinkedIn users on darknet', The Guardian. Retrieved from: https://www.theguardian.com/technology/2016/may/18/hacker-advertises-details-of-117-million-linkedin-users-on-darknet – last accessed October 15 2018.

Thielman, S (December 15, 2016), 'Yahoo hack: 1bn accounts compromised by biggtest data breach in history', The Guardian. Retrieved from: https://www.theguardian.com/technology/2016/dec/14/yahoo-hack-security-of-one-billion-accounts-breached – last accessed October 15 2018.

Hunt, T (March 29, 2018), 'Have I Been Pwned is Now Partnering With 1Password'. Retrieved from: https://www.troyhunt.com/have-i-been-pwned-is-now-partnering-with-1password/ – last accessed October 15 2018.

Wong, J C (November 22, 2017), 'Uber concealed massive hack that exposed data of 57m users and drivers', The Guardian. Retrieved from: https://www.theguardian.com/technology/2017/nov/21/uber-data-hack-cyber-attack – last accessed October 15 2018.

[1] Researcher for public law, constitutional law and data protection law regarding technical aspects at FIZ Karlsruhe, Leibniz Institute for Information Infrastructure. The author thanks Dr. Dara Hallinan for his helpful advices during the writing process.

[2] Wong (November 22 2017), 'Uber concealed massive hack that exposed data of 57m users and drivers', The Guardian. Retrieved from: https://www.theguardian.com/technology/2017/nov/21/uber-data-hack-cyber-attack – last accessed October 15 2018.

[3] Thielman, S (December 15 2016), 'Yahoo hack: 1bn accounts compromised by biggest data breach in history', The Guardian. Retrieved from: https://www.theguardian.com/technology/2016/dec/14/yahoo-hack-security-of-one-billion-accounts-breached – last accessed October 15 2018.

[4] Press Association (May 18 2016), 'Hacker advertises details of 117 million LinkedIn users on darknet', The Guardian. Retrieved from: https://www.theguardian.com/technology/2016/may/18/hacker-advertises-details-of-117-million-linkedin-users-on-darknet – last accessed October 15 2018.

[5] A brief overview of websites with this feature is available at https://www.twofactorauth.org – last accessed October 15 2018.

[6] See https://www.haveibeenpwned.com – last accessed October 15 2018.

[7] The mention of 1Password is reasoned in the cooperation of both, so 1Password uses the leak checking tool for passwords and email addresses. More information in Hunt, T (March 29 2018), 'Have I Been Pwned is Now Partnering With 1Password' and 1Password (February 22, 2018), 'Finding Pwned Passwords with 1Password'.

[8] https://breachalarm.com – last accessed October 15 2018.

[9] http://hacked-emails.com – last accessed October 15 2018.

[10] http://leakedsource.ru – last accessed October 15 2018.

[11] http://leakbase.pw – last accessed October 15 2018.

[12] https://sec.hpi.de/ilc/search – last accessed October 15 2018.

[13] Retrievable at https://www.sicherheitstest.bsi.de – last accessed October 15 2018.

[14] Further explanations can be found in the German press release from the German Federal Office for Information Security. See BSI (January 21, 2014), 'Millionenfacher Identitätsdiebstahl: BSI bietet Sicherheitstest für E-Mail-Adressen'. Retrieved from: https://www.bsi.bund.de/DE/Presse/Pressemitteilungen/Presse2014/Mailtest_21012014.html – last accessed October 15 2018.

[15] See the twelve recommendations for self-made data protection in BSI, 'Zwölf Maßnahmen zur Absicherung gegen Angriffe aus dem Internet'. Retrieved from: https://www.bsi-fuer-buerger.de/BSIFB/DE/Service/Checklisten/Massnahmen_gegen_Internetangriffe.html – last accessed October 15 2018.

[16] For further information see https://www.melani.admin.ch/melani/en/home.html – last accessed October 15 2018.

[17] See https://www.antiphishing.ch/en/ – last accessed October 15 2018.

[18] https://www.melani.admin.ch/melani/en/home/meldeformular/use.html – last accessed October 15 2018.

[19] https://checktool.ch – last accessed October 15 2018.

[20] https://swissleak.ch – last accessed October 15 2018.

[21] Just to explain this problem shortly: Switzerland is not a Member State of the European Union, why the GDPR isn't regulating the data protection law in Switzerland directly. But this doesn't mean that the GDPR fails in regulating the use of personal data across the border: On one hand, the special regulation for personal data in case of Art. 3(2)(a) GDPR will still be applicable. So, e.g. if the controller is processing personal data of a data subject from the EU and when a service is delivered "across the border of Switzerland", the controller must consider Art. 33, 34 GDPR. In turn, personal data from data subjects in Switzerland isn't regulated regarding the processing by Swiss companies. This

situation is only ruled by national law. On the other hand, the bilateral treaties between Switzerland and the EU have to be considered, which may contain a basis for a data protection treaty. Besides that, there is still the rumor that the existing level of data protection in Switzerland will be adjusted to agree with the GDPR. Since early 2017, the Federal Council is working on an amendment for the DSG.

[22] See the German ranking in 2016 at https://www.heise.de/security/meldung/hallo-ist-meistgenutztes-deutsches-Passwort-auf-Platz-zehn-steht-ficken-3579567.html and the American ranking in 2017 http://time.com/5071176/worst-passwords-2017/.

[23] *Patrick Breyer v Bundesrepublik Deutschland* [2016] ECJ C-582/14, paragraph 47.

[24] Bäcker in Kühling, J and Buchner, B (2017), Art. 14 DSGVO, paragraph 55.

[25] See Data Protection Act 2018, chapter 12, s. 8. In the German BDSG, similar reasons are distributed in several Paragraphs, e.g. § 24 and 28 BDSG.

[26] Albrecht, J P and Jotzo F (2017), Das neue Datenschutzrecht der EU, pp. 72 f. Similar Buchner, B and Petri, T in Kühling, J and Buchner, B (2017), Art. 6 DSGVO, paragraph 76 f.

[27] Buchner, B and Petri, T in Kühling, J and Buchner, B (2017), DSGVO, Art. 6, paragraph 142; Robrahn R and Bremert B (2017), pp. 291-293.

[28] Robrahn R and Bremert B (2017), page 923.

[29] *Patrick Breyer v Bundesrepublik Deutschland* [2016] ECJ C-582/14, paragraph 47.

[30] GDPR, Recital 49.

[31] Hartung in Kühling, J and Buchner, B (2017), Art. 25 DSGVO, paragraph 16.

[32] A more detailed explanation can be found at https://www.youtube.com/watch?v=HUs1bH85X9I – last accessed October 15 2018.

[33] For more information see the German BSI catalogue for IT-security, available at https://www.bsi.bund.de/DE/Themen/ITGrundschutz/ITGrundschutzKataloge/itgrundschutzkataloge_node.html – last accessed October 15 2018.

[34] Similar *Patrick Breyer v Bundesrepublik Deutschland* [2016] ECJ C-582/14, paragraph 47 and German Federal Constitutional Court decision no. 1 BvR 209/83, pp. 45: “By (automated) aggregation or combination, irrelevant data can get a new worth.” – translated by the author. Regarding different kinds of pseudonyms see Rossnagel A and Scholz P (2000), pp. 723, 725 f, 727.

[35] In Art. 1 (1) of the Council Directive 91/250/EEC (Computer Program Directive), the term “computer program” is explained in a literary way, so only the source code and possibly also including the design material. This distinction is also mentioned in Art. 1 (3) and 2 Council Directive 96/9/EC. See also Vohwinkel, M in Ahlberg, H and Götting, H-P (20.04.2018), § 87a UrhG, paragraph 10 f.

[36] Vohwinkel, M in Ahlberg, H and Götting, H-P (20.04.2018), § 87a UrhG, paragraph 31 f.