Information Modelling and Knowledge Bases XXXII
M. Tropmann-Frick et al. (Eds.)

doi:10.3233/FAIA200831

Automated Text Similarities Approach: GDPR and Privacy by Design Principles

Boštjan BRUMEN
University of Maribor (www.um.si), Faculty of Electrical Engineering and Computer science, Smetanova 17, Si-2000 Maribor, Slovenia
bostjan.brumen@uni-mb.si

Abstract. Respect for privacy is not a modern phenomenon as it has been around for centuries. Recent advances in technologies led to the rise of awareness of the importance of privacy, and to the development of principles for privacy protection to guide the engineering of information systems on one side, and on using the principles to draft legal texts protecting privacy on the other side. In this paper, we analyze how respect for privacy has been implemented in GDPR by automated comparison of the similarity of GDPR's articles and the text of seven principles of Privacy by Design. We have compared the specific text of GDPR's first 50 core privacy-protecting articles and the GDPR's remaining provisions to establish independent supervisory authorities. The first half is observing the privacy by design principles, each of them considerably more than the second half. Our findings show that automated similarity comparison can highlight portions of legal texts where principles were observed. The results can support drafting legal texts to check whether important legal (or other) principles were adequately addressed.

Keywords. Privacy, GDPR, information system, privacy by design, similarity, semantics

1. Introduction

Privacy is not a new phenomenon. The existence of two areas, the public area of politics and political activity, the *polis* (gr. πολις), and the private one of the family, the *oikos* (gr. οίκος), as two interdependent and sometimes conflicting areas, was well known in the times of Ancient Greek civilization [1, 2], and was reflected in classic dramas, e.g., in Sophocles' Antigone and Oedipus Rex. Interestingly, the New Order of the *polis*, despite its presumed weaknesses, reigns supreme at the end of the dramas [3].

Privacy was an essential issue in the medical profession as well: *"...Whatever I see or hear in the lives of my patients, whether in connection with my professional practice or not, which ought not to be spoken of outside, I will keep secret, as considering all such things to be private. ..."* [4] is the text from Hippocrates' oath that addresses privacy and instructs ancient doctors to keep private data – secret! Privacy and confidentiality are significant contemporary issues, especially in the Western world, and are not limited to the medical field only [5].

Privacy has re-emerged as a vital issue post the widespread use of the Internet and world wide web, a new ecosystem for data with all new challenges. The emergence of social networks has worsened the protection of private data. What users thought would remain private could and actually was used against their will and/or consent.

In light of preparations for European Union's new regulations on data protection in 2013, Mark Zuckerberg, the founder of the most used social network Facebook, and Facebook's chief operation officer Sheryl Sandberg stated that the privacy controls were centered at Facebook's core at all times [6, 7].

Then happened the Cambridge Analytica. Between 50 and 87 million Facebook user profiles, depending on the source ([8] and [9], respectively), were collected in a manner that users neither foresaw nor allowed. Previously, volunteers were analyzed using the "OCEAN" psychological profile (openness, conscientiousness, extraversion, agreeableness, and neuroticism) and correlated it with their Facebook activity (likes and shares), demonstrating that Facebook profile data could be used instead of a formal psychographic instrument [9]. It then used the test results and Facebook data to build an algorithm that could analyze individual Facebook profiles and determine personality traits linked to voting behavior [8]. Fifty million profiles at the time represented around a third of active North American Facebook users and nearly a quarter of potential U.S. voters [8]. Displaying individualized, high impact messages to swing voters is sufficient to impact election results in a few states, especially in small ones with as few as a couple of hundred thousand voters [9, 10].

Facebook denied that the harvesting of millions of profiles by Cambridge Analytica was a data breach and hence failed to report the regulators and individuals about the breach[8].

After two weeks, the Cambridge Analytica scandal broke out, Facebook via Mark Zuckerberg apologized for a "breach of trust" in several U.S. and U.K. newspapers adds [11]: "*I'm sorry we didn't do more at the time. We're now taking steps to ensure this doesn't happen again.*"

Firstly, it was not only a breach of trust; it was a breach of privacy. Secondly, based on previous experiences, we can rest assured it will happen again.

Jim Isaak and Mina J. Hanna wrote: "*It is clear that national governance institutions demonstrably lack the ability to anticipate technology's future impact on the rights and duties of its citizens, much less its impact on the structure of society, ideological divides, and political schisms among its citizens and the expansion of identity politics promoted by isolated social and news media echo chambers.*"

The Cambridge Analytica scandal has firstly shown that there are many databases containing private data, and they are readily available to be bought or exploited. Secondly, the microtargeting of individuals is doable not only illegally, but (currently) also legally, without disclosure and informed consent, completely bypassing laws and regulations. Thirdly, the expenses for doing it are negligible and yields at high stakes. Lastly, corporations storing and processing the data are rarely held responsible and fined appropriately.

All this calls for changes in corporate and government levels. Corporations should anticipate legal changes, and governments must ensure that private and/or personal data are protected so that individuals can best exercise their citizens' statutory and constitutional rights, such as due process, equal representation before the law, the right to appeal, freedom of expression, voting, and non-discrimination [9].

The laws of most developed countries impose obligations to respect informational privacy (e.g., confidentiality, anonymity, secrecy, and data security); physical privacy (e.g., modesty and bodily integrity); associational privacy (e.g., intimate sharing of death, illness, and recovery); proprietary privacy (e.g., self-ownership and control over personal identifiers, genetic data, and body tissues); and decisional privacy (e.g., autonomy and choice in personal relationships) [12, 13]. It is the lack of respect for the laws or the

incompleteness of these laws that privacy is not protected adequately. The lack of respect comes from either ignorance or deliberation. While deliberate acts will always happen – and need to be sanctioned appropriately –, the ignorance and incompleteness must be addressed.

One way to address the issues is by following the already established principles on protecting privacy and start protecting at the beginning of the processes – by following the privacy by design principles.

In this paper, the research question is how European regulation on data processing, the famous General Data Protection Regulation (GDPR) directive [14] – the Regulation (E.U.) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data – is addressing the privacy by using the privacy by design principles.

The rest of the paper is organized as follows. In Section 2, we give an overview of the literature review dealing with privacy and privacy by design principle. In Section 3, we describe our research method and present the results. In Section 4, we conclude the paper with final remarks.

2. Literature review

"*Privacy is a concept in disarray. Nobody can articulate what it means*." is an interesting observation by D. Solove [15]. Nevertheless, privacy and confidentiality were well researched and addressed by philosophers and jurists alike and later addressed by many technologies.

A systematic discussion on the concept of privacy has begun with the famous article by Samuel Warren and Louis Brandeis titled "The Right to Privacy" [16]. Citing "*political, social, and economic changes*" and a recognition of "*the right to be let alone,*" they argued that existing law (i.e., the Constitution of the U.S.A.) afforded a way to protect the privacy of the individual, and they sought to explain the nature and extent of that protection. Focusing in large part on the press and publicity allowed by then "*recent*" inventions such as photography and newspapers, but referring as well to violations in other contexts, they emphasized the invasion of privacy brought about by public dissemination of details relating to a person's private life. Warren and Brandeis felt a variety of existing cases could be protected under a more general right to privacy, which would protect the extent to which one's thoughts, sentiments, and emotions could be shared with others. Urging that they were not attempting to protect the items produced, or intellectual property, but rather the peace of mind attained with such protection, they said the right to privacy was based on a principle of "inviolate personality," which was part of a general right of immunity of the person, "the right to one's personality" [16]. Warren and Brandeis thus laid the legal foundation for a concept of privacy that has come to be known as control over information about oneself [17].

In 1960, Prosser systematically defined four different aspects of "*privacy rights*" being upheld in tort law: [17, 18]:

1. Intrusion upon a person's seclusion or solitude, or into his private affairs.
2. Public disclosure of embarrassing private facts about an individual.
3. Publicity placing one in a false light in the public eye.
4. Appropriation of one's likeness for the advantage of another [17].

Prosser noted that the intrusion in the first privacy right had expanded beyond physical intrusion and pointed out that Warren and Brandeis had been concerned primarily with the second privacy right. Nevertheless, Prosser felt that both real abuses and public demand had led to the general acceptance of these four types of privacy invasions. Thomas Nagel, one of America's top contemporary philosophers, gives a more contemporary (philosophical) discussion of privacy, concealment, publicity, and exposure [17, 19].

As summarized by authors in [5], Adam Moore [20], building on the views of Ruth Gavison [21], Anita Allen [22], Sissela Bok [23], and others, offers a "control over access" account of privacy. According to Moore, privacy is a culturally and species relative right to a level of control over access to bodies or places and information. While defending the view that privacy is relative to species and culture, Moore argues that privacy is objectively valuable: human beings that do not obtain a certain level of control over access will suffer in various ways. Moore claims that privacy, like education, health, and maintaining social relationships, is an essential part of human flourishing or well-being [17].

In medical contexts, as viewed by Allen [13], the "privacy" at issue is very often "confidentiality" [24], specifically the confidentiality of patient-provider encounters (including the very fact that an encounter has taken place), along with the secrecy and security of information memorialized in physical, electronic and graphic records created as a consequence of patient-provider encounters [24]. Confidentiality is defined as restricting information to persons belonging to a set of specifically authorized recipients [13, 22, 25, 26]. Confidentiality can be achieved either through professional silence, leaning on the moral aspect, or through secure data management [27], leaning on technologies and techniques.

The moral significance attached to privacy is reflected in data protection and security regulations adopted by local and national authorities around the world. One such regulation is the European Union's General Data Protection Regulation.

2.1. Privacy by design

The literature presented above has shown that privacy has to be taken seriously as it addresses one of the fundamental human rights and has a special place in legal texts. It is explicitly stated under Article 12 of the 1948 Universal Declaration of Human Rights and protected by 1st, 3rd, 4th, and 5th Amendment of the U.S. Constitution [15]. In European Union, it is protected by Article 8(1) of the Charter of Fundamental Rights of the European Union and Article 16(1) of the Treaty on the Functioning of the European Union (TFEU), and several national constitutions [15].

The Ontario Privacy Commissioner Ann Cavoukian has developed a "Privacy by Design" (PbD) framework [28-31], which is emphasizing the need to adopt a proactive rather than a reactive compliance approach to the protection of privacy. To safeguard privacy, legislation and regulation would no longer be sufficient; privacy needs to be proactively embedded directly into information technology, business practices, physical design, and networked infrastructures – making it the default [32]. Interestingly, the framework can also be applied when designing legal procedures [33, 34] and has become an international standard for assuring privacy in the information era [32].

The framework relies on 7 principles [28], see Figure 1:

1. Proactive not Reactive; Preventative not Remedial

The meaning of the principle reads: "*The Privacy by Design approach is characterized by proactive rather than reactive measures. It anticipates and prevents privacy invasive events before they happen. Privacy by design does not wait for privacy risks to materialize, nor does it offer remedies for resolving privacy infractions once they have occurred — it aims to prevent them from occurring. In short, Privacy by Design comes before-the-fact, not after.*" [28]

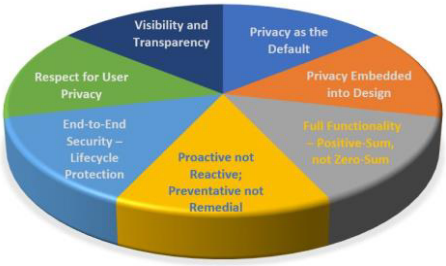

Figure 1: Seven principles of Privacy by Design

2. Privacy as the Default

"Privacy as the Default" principle reads: "*We can all be certain of one thing — the default rules! Privacy by design seeks to deliver the maximum degree of privacy by ensuring that personal data are automatically protected in any given I.T. system or business practice. If an individual does nothing, their privacy still remains intact. No action is required on the part of the individual to protect their privacy — it is built into the system, by default.*" [28]

3. Privacy Embedded into Design

"*Privacy by Design is embedded into the design and architecture of I.T. systems and business practices. It is not bolted on as an add-on, after the fact. The result is that privacy becomes an essential component of the core functionality being delivered. Privacy is integral to the system, without diminishing functionality.*" [28]

4. Full Functionality – Positive-Sum, not Zero-Sum

"*Privacy by Design seeks to accommodate all legitimate interests and objectives in a positive-sum win-win manner, not through a dated, zero-sum approach, where unnecessary trade-offs are made. Privacy by design avoids the pretense of false dichotomies, such as privacy vs. security – demonstrating that it is possible to have both.*" [28]

5. End-to-End Security – Lifecycle Protection

"*Privacy by Design, having been embedded into the system prior to the first element of information being collected, extends securely throughout the entire lifecycle of the data involved — strong security measures are essential to privacy, from start to finish. This ensures that all data are securely retained, and then securely destroyed at the end of the process, in a timely fashion. Thus, Privacy by Design ensures cradle to grave, secure lifecycle management of information, end-to-end.*" [28]

6. Visibility and Transparency

"*Privacy by Design seeks to assure all stakeholders that whatever the business practice or technology involved, it is in fact, operating according to the stated promises and objectives, subject to independent verification. Its component parts and operations remain visible and transparent, to users and providers alike. Remember, trust but verify*." [28]

7. Respect for User Privacy

"*Above all, Privacy by Design requires architects and operators to protect the interests of the individual by offering such measures as strong privacy defaults, appropriate notice, and empowering user-friendly options. Keep it user-centric*." [28]

The European GDPR was drafted with Privacy by Design as one of the guiding frameworks [35, 36]. It is to notice that Article 25 of GDPR is titled "Data protection by design and by default", and that Recital 78 is mentioning the principles of data protection by design and by default.

In the next section, we will analyze the GDPR and answer the question, to what extent is GDPR, throughout its articles, addressing each of the 7 principles.

3. Analysis of GDPR and seven Privacy by Design principles

In this section, we check how each of the principles is reflected or addressed in GDPR.

The text of GDPR is naturally divided into two major parts. The first part consists of general provisions (Chapter I), principles (Chapter II), rights of the data subject (Chapter III), controller and processor (Chapter IV), and transfers of personal data to third countries or international organizations (Chapter V). The second part consists of articled defining the independent supervisory authorities (Chapter VI), cooperation and consistency (Chapter VII), remedies, liability and penalties (Chapter VIII), provisions relating to specific processing situations (Chapter IX), delegated acts, and implementing acts (Chapter X), and final provisions (Chapter XI).

In the first half, the first fifty articles represent the core of the GDPR and its intent to protect privacy. In the second half, the remaining articles (§51-§96) are provisions for establishing independent supervisory authorities (e.g., privacy commissioners) and for remedies, liability, and penalties, following by concluding articles. Hence, we separately studied the first and the second half of the GDPR and its relative semantic similarity to seven Privacy by Design principles.

We measure the extent to which these principles are reflected in GDPR's articles by using the automated text similarities approach and the Universal Sentence Encoder (USE) [37]. USE is a pre-trained sentence encoder which encodes text paragraphs into high dimensional vectors that can be used for detecting semantic similarity (and other natural language tasks, such as text classification or clustering), see Figure 2. USE is typically pre-trained on a range of supervised and unsupervised tasks in order to comprehend semantic information in texts [38]. It is learning from various data sources and on diverse tasks to dynamically accommodate a wide variety of natural language understanding tasks [37].

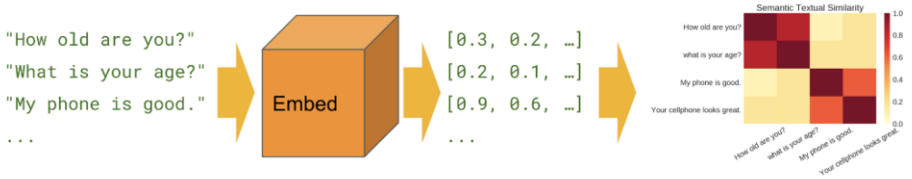

Figure 2: Process of embedding and comparing of different texts [38]

The input is variable-length English text, and the output is a 512-dimensional vector. The embeddings produced by the USE are approximately normalized. The semantic similarity of two sentences can be trivially computed as the inner product of the encodings.

We used the Google's Semantic Similarity with TF-Hub Universal Encoder online tool, available at https://colab.research.google.com/github/tensorflow/hub/blob/master/examples/colab/semantic_similarity_with_tf_hub_universal_encoder.ipynb, and calculated the matrix of inner products between encodings of 7 principles' text, and encodings of each individual GDPR article's text. In the latter, we only removed numberings of paragraphs or sections (e.g., "1." and "(a)" were removed). Each article's text was joined into a single paragraph to be able to process it.

The automatically calculated semantic similarity was checked against the S.T.S. Benchmark [39] (http://ixa2.si.ehu.es/stswiki/index.php/STSbenchmark). It evaluates the degree to which similarity scores computed using USE are in line with the human evaluation of similarity. The benchmark uses similarity scores for a diverse selection of sentence pairs. Pearson's R is calculated to estimate the relationship (quality) between the automatically generated similarity scores and human evaluations. There is a strong, positive correlation between machine similarity scores and human evaluations, which is statistically significant ($R = 0.803$, $p < 0.005$).

The similarity scores between the GDPR's first 50 articles and seven principles of Privacy by Design are presented in Table 1. A value of 0 represents no similarity, and value 1 represents perfect semantic similarity (equal texts).

Not surprisingly, the most similarity was found between the 2nd principle (Privacy as the Default) and GDPR's articles. 2nd principle scored 32 maximum values (blue shaded cells of Table 1), followed by 6th principle (Visibility and Transparency) with 13, 1st principle (Proactive not Reactive; Preventative not Remedial) with 3 and 5th principle (End-to-End Security – Lifecycle Protection) with 2 maximum values.

Principles with the most maximum values also had the least minimum values. The least similarity was found between the 7th principle (Respect for User Privacy) with 16 minimum values (red shaded cells), followed by 3rd, 4th, 6th, 1st and 5th principle with 12, 11, 7, 2 and 2 minimum values, respectively.

Principle 2 was not only in general most like GDPR's text, it was also most strongly similar. It had largest similarity index for articles 1, 34, 21, 33, 5, 14, 25, 18, 20, 15, 17, 11, 13, 7, 22, 16, 4, 9 and 50, similarity values ranging from 0.491 to 0.326.

Article 5 was most similar to *all* 7 principles with average similarity value of 0.375, and Article 31 was least similar with average similarity value of 0.054. Thus, automatic semantic encoder correctly identified the Article 5 that lists principles of GDPR; it's title is "*Principles relating to processing of personal data*". Article 31's title is "Cooperation with the supervisory authority", requiring controller and processor to cooperate with the supervisory body; the material not covered by any of the seven principles, hence article's average similarity is extremely low.

Table 1: Similarity scores between GDPR's articles 1-50 and 7 Privacy by Design principles

Article	Principle 1	Principle 2	Principle 3	Principle 4	Principle 5	Principle 6	Principle 7	Max	Min	Average
1	0.360347	0.491606	0.289023	0.338619	0.284758	0.260049	0.347171	0.491606	0.260049	0.338796
2	0.236012	0.303697	0.234517	0.185280	0.238605	0.240943	0.183850	0.303697	0.183850	0.231843
3	0.199190	0.275385	0.211665	0.156431	0.201384	0.288125	0.098190	0.288125	0.098190	0.204339
4	0.277767	0.328543	0.234926	0.249763	0.255424	0.265395	0.183510	0.328543	0.183510	0.256475
5	0.389129	0.432845	0.297043	0.425751	0.348858	0.344788	0.387155	0.432845	0.297043	0.375081
6	0.254904	0.320956	0.217383	0.257224	0.251671	0.245881	0.224717	0.320956	0.217383	0.253248
7	0.229825	0.350499	0.148846	0.216655	0.238838	0.234083	0.195935	0.350499	0.148846	0.230669
8	0.206144	0.259325	0.103134	0.144194	0.133171	0.236139	0.095653	0.259325	0.095653	0.168252
9	0.299966	0.328485	0.218131	0.273811	0.221169	0.219755	0.240247	0.328485	0.218131	0.257366
10	0.288422	0.319372	0.196702	0.204103	0.249750	0.253504	0.194260	0.319372	0.194260	0.243731
11	0.304790	0.368847	0.225721	0.227445	0.224513	0.250568	0.163251	0.368847	0.163251	0.252162
12	0.221201	0.237019	0.104119	0.193814	0.166171	0.163481	0.148146	0.237019	0.104119	0.176279
13	0.253057	0.362358	0.211481	0.254764	0.259486	0.198734	0.165719	0.362358	0.165719	0.243657
14	0.319966	0.400297	0.253912	0.297740	0.284296	0.238595	0.212792	0.400297	0.212792	0.286800
15	0.238152	0.373431	0.185020	0.223043	0.235669	0.180948	0.188269	0.373431	0.180948	0.232076
16	0.140253	0.337666	0.146838	0.187919	0.245466	0.119677	0.269672	0.337666	0.119677	0.206785
17	0.307518	0.370307	0.267838	0.322436	0.247481	0.220210	0.207319	0.370307	0.207319	0.277587
18	0.303190	0.385684	0.236388	0.318092	0.247065	0.210709	0.227209	0.385684	0.210709	0.275477
19	0.241875	0.324473	0.152771	0.225434	0.172970	0.271217	0.191730	0.324473	0.152771	0.225781
20	0.234367	0.374052	0.223007	0.231875	0.251366	0.143427	0.161639	0.374052	0.143427	0.231391
21	0.313663	0.439330	0.268501	0.315681	0.316107	0.252327	0.242396	0.439330	0.242396	0.306858
22	0.278818	0.347352	0.223850	0.233086	0.242484	0.211744	0.215365	0.347352	0.211744	0.250385
23	0.271151	0.253360	0.194426	0.246279	0.231250	0.225641	0.226124	0.271151	0.194426	0.235462
24	0.125837	0.083455	0.062519	0.113094	0.103950	0.166676	0.163227	0.166676	0.062519	0.116965
25	0.319180	0.389360	0.307548	0.295063	0.376919	0.324592	0.323794	0.389360	0.295063	0.333779
26	0.079373	0.197641	0.124196	0.082077	0.142194	0.249562	0.178711	0.249562	0.079373	0.150536
27	0.221496	0.243336	0.202414	0.174030	0.149802	0.234979	0.151558	0.243336	0.149802	0.196802
28	0.091010	0.121777	0.151472	0.080631	0.136006	0.210217	0.068764	0.210217	0.068764	0.122840
29	0.074639	0.215947	0.202555	0.050227	0.151180	0.242613	0.090598	0.242613	0.050227	0.146823
30	0.108839	0.185647	0.158982	0.084969	0.129354	0.201238	0.087208	0.201238	0.084969	0.136605
31	0.000960	0.040138	0.003260	0.028368	0.023398	0.156040	0.125144	0.156040	0.000960	0.053901
32	0.280560	0.306986	0.265131	0.256895	0.326574	0.310465	0.236674	0.326574	0.236674	0.283327
33	0.428107	0.435881	0.266004	0.276108	0.322867	0.274054	0.285993	0.435881	0.266004	0.327002
34	0.376568	0.466291	0.245363	0.322677	0.327622	0.284251	0.320097	0.466291	0.245363	0.334695
35	0.291783	0.305647	0.205283	0.186971	0.279808	0.270116	0.260022	0.305647	0.186971	0.257090
36	0.282082	0.237424	0.176635	0.142290	0.241672	0.233886	0.203028	0.282082	0.142290	0.216717
37	0.175927	0.274563	0.204100	0.124045	0.227051	0.270106	0.228610	0.274563	0.124045	0.214914
38	0.245430	0.326263	0.212728	0.197173	0.309278	0.336880	0.283499	0.336880	0.197173	0.273036
39	0.144685	0.148925	0.095902	0.090626	0.172297	0.224215	0.164543	0.224215	0.090626	0.148742
40	0.170306	0.145069	0.080152	0.081939	0.112100	0.164508	0.140494	0.170306	0.080152	0.127795
41	0.204474	0.181865	0.193979	0.157812	0.166187	0.258832	0.146790	0.258832	0.146790	0.187134
42	0.102236	0.086630	0.146038	0.016430	0.136060	0.214091	0.041492	0.214091	0.016430	0.106140
43	0.087727	0.033069	0.097247	0.010421	0.075268	0.128426	0.007185	0.128426	0.007185	0.062763
44	0.099720	0.208492	0.127479	0.086833	0.127452	0.121329	0.087357	0.208492	0.086833	0.122666
45	0.153965	0.163503	0.136026	0.081461	0.150335	0.139639	0.078758	0.163503	0.078758	0.129098
46	0.084423	0.115119	0.109503	0.039761	0.162764	0.085834	0.073094	0.162764	0.039761	0.095785
47	0.234158	0.256798	0.183733	0.187696	0.240033	0.349676	0.202147	0.349676	0.183733	0.236320
48	0.015412	0.123173	0.082316	0.079838	0.003679	0.073136	0.018383	0.123173	0.003679	0.056562
49	0.218997	0.270453	0.177729	0.213976	0.202159	0.139674	0.120007	0.270453	0.120007	0.191856
50	0.251524	0.326452	0.159450	0.186421	0.200024	0.100079	0.272655	0.326452	0.100079	0.213801
Average	0.220783	0.277496	0.184460	0.187545	0.210880	0.220821	0.182603			
Max	0.428107	0.491606	0.307548	0.425751	0.376919	0.349676	0.387155			
Min	0.000960	0.033069	0.003260	0.010421	0.003679	0.073136	0.007185			

Similarities between GDPR's articles 1-50 and 7 principles can easily be seen in Figure 3. Most similarities are found between GDPR and 2nd principle; 1st and 5th are quite similar too, especially in the first part of GDPR, up to article 25.

Additionally, more similarity is found between all principles and Articles 32-38. These articles deal with security of personal data an belong to Chapter IV titled "Controller and processor" [of personal data], Sections 2-4 titled "Security of personal data", "Data protection impact assessment and prior consultation" and "Data protection officer", respectively.

A low similarity can be found in Article 24, and 26-31. Article 24 deals with responsibilities of the controller, which is a general legal text. Articles 26-31 deal with joint controllers and processors of data.

There is relatively low similarity between 7 principles and GDPR's Articles from 38 onwards. These articles deal with tasks of the data protection officer (Article 39), with "Codes of conduct and certification" (Section 5 of Chapter IV, Articles 40-43) and with "Transfers of personal data to third countries or international organisations" (Chapter V, Articles 44-50).

The second half of the GDPR's articles (§51-§96) and the similarity scores between them and 7 principles of Privacy by Design are presented in Table 2.

Table 2: Similarity scores between GDPR's articles 51-96 and 7 Privacy by Design principles

Article	Principle 1	Principle 2	Principle 3	Principle 4	Principle 5	Principle 6	Principle 7	Max	Min	Average
51	0,143308	0,192724	0,085728	0,022624	0,200529	0,159537	0,131606	0,200529	0,022624	0,133722
52	0,188505	0,262597	0,159334	0,081576	0,194207	0,389725	0,142407	0,389725	0,081576	0,202622
53	0,093270	0,194384	0,104460	0,047490	0,165538	0,244022	0,095565	0,244022	0,047490	0,134961
54	0,167010	0,231476	0,102635	0,078626	0,171476	0,267520	0,150630	0,267520	0,078626	0,167053
55	0,107071	0,175314	0,071167	0,018215	0,070886	0,144848	0,057889	0,175314	0,018215	0,092199
56	0,156473	0,136461	0,095025	0,044338	0,165821	0,189976	0,057899	0,189976	0,044338	0,120856
57	0,194033	0,156610	0,099421	0,116906	0,144530	0,193734	0,168287	0,194033	0,099421	0,153360
58	0,186682	0,171167	0,121743	0,086001	0,129005	0,207121	0,112626	0,207121	0,086001	0,144906
59	0,131349	0,158286	0,026601	0,011019	0,067565	0,092920	0,138071	0,158286	0,011019	0,089402
60	0,210469	0,120840	0,082329	0,113166	0,160432	0,172535	0,131482	0,210469	0,082329	0,141608
61	0,202218	0,182065	0,070386	0,141082	0,098679	0,125516	0,167967	0,202218	0,070386	0,141130
62	0,136179	0,149914	0,085658	0,030501	0,099688	0,189076	0,066159	0,189076	0,030501	0,108168
63	0,083103	0,026745	0,018116	0,077494	0,101798	0,068248	0,040582	0,101798	0,018116	0,059441
64	0,147587	0,137618	0,093579	0,097216	0,144813	0,099472	0,063285	0,147587	0,063285	0,111939
65	0,155961	0,091938	0,062470	0,079194	0,135657	0,111003	0,020748	0,155961	0,020748	0,093853
66	0,255523	0,189699	0,113062	0,205661	0,144996	0,095848	0,169913	0,255523	0,095848	0,167815
67	0,058775	0,072375	0,024739	0,006772	0,061054	0,043697	0,039761	0,072375	0,006772	0,043882
68	0,107859	0,202085	0,123124	0,030473	0,146351	0,161756	0,071877	0,202085	0,030473	0,120504
69	0,154528	0,097043	0,046815	0,106469	0,000858	0,164906	0,021910	0,164906	0,000858	0,084647
70	0,160756	0,149889	0,091806	0,090808	0,167566	0,099980	0,172899	0,172899	0,090808	0,133386
71	0,107687	0,103234	0,059864	0,065568	0,064584	0,020711	0,125060	0,125060	0,020711	0,078101
72	0,145450	0,261305	0,177547	0,087010	0,165080	0,224601	0,183813	0,261305	0,087010	0,177829
73	0,045818	0,048667	0,077140	0,051846	0,042214	0,101706	0,036437	0,101706	0,036437	0,057690
74	0,072820	0,020418	0,036130	0,019769	0,054040	0,051676	0,017552	0,072820	0,017552	0,038915
75	0,069851	0,139503	0,131429	0,030529	0,148549	0,073375	0,088830	0,148549	0,030529	0,097438
76	0,147655	0,274865	0,256131	0,152684	0,146188	0,195388	0,142149	0,274865	0,142149	0,187866
77	0,289939	0,288280	0,165043	0,232277	0,132183	0,096584	0,189712	0,289939	0,096584	0,199145
78	0,234765	0,183912	0,093571	0,139857	0,172520	0,059264	0,050159	0,234765	0,050159	0,133435
79	0,219171	0,305313	0,187409	0,176196	0,192825	0,133416	0,182733	0,305313	0,133416	0,199580
80	0,224790	0,269500	0,130137	0,193417	0,174397	0,158391	0,163115	0,269500	0,130137	0,187678
81	0,121552	0,089024	0,045705	0,031475	0,118749	0,095484	0,071226	0,121552	0,031475	0,081888
82	0,054833	0,144331	0,087929	0,048178	0,082467	0,132232	0,077667	0,144331	0,048178	0,089662
83	0,138593	0,177329	0,113523	0,059127	0,148548	0,151006	0,109593	0,177329	0,059127	0,128246
84	0,093019	0,163261	0,016305	0,015677	0,049928	0,033648	0,135812	0,163261	0,015677	0,072521
85	0,223475	0,250794	0,172253	0,192507	0,152005	0,190728	0,203694	0,250794	0,152005	0,197922
86	0,260916	0,194505	0,115236	0,275363	0,160502	0,196304	0,165471	0,275363	0,115236	0,195471
87	0,355063	0,379704	0,290070	0,283847	0,190952	0,285546	0,274608	0,379704	0,190952	0,294256
88	0,145747	0,150947	0,038417	0,111588	0,061154	0,142636	0,126429	0,150947	0,038417	0,110988
89	0,075127	0,017228	0,028606	0,001717	0,011263	0,034014	0,063407	0,075127	0,001717	0,033052
90	0,010111	0,099717	0,104467	0,020106	0,062799	0,062357	0,049598	0,104467	0,010111	0,058451
91	0,141641	0,180350	0,117856	0,112441	0,178418	0,084871	0,075489	0,180350	0,075489	0,127295
92	0,137399	0,201454	0,136692	0,224414	0,114731	0,236259	0,198356	0,236259	0,114731	0,178472
93	0,059384	0,059658	0,041880	0,001147	0,089387	0,115351	0,023873	0,115351	0,001147	0,055811
94	0,083752	0,109499	0,058559	0,038162	0,079972	0,066910	0,064770	0,109499	0,038162	0,071661
95	0,188060	0,290234	0,100327	0,178483	0,209332	0,174625	0,329203	0,329203	0,100327	0,210038
96	0,039346	0,056843	0,022354	0,032676	0,009920	0,013272	0,052202	0,056843	0,009920	0,032373
Average	0,146231	0,164328	0,097452	0,092645	0,121395	0,138083	0,113533			
Max	0,355063	0,379704	0,29007	0,283847	0,209332	0,389725	0,329203			
Min	0,010111	0,017228	0,016305	0,001147	0,000858	0,013272	0,017552			

In the second part of GDPR the most similarity was again found between 2nd principle (Privacy as the Default) and GDPR's articles. 2nd principle scored 17 maximum

values (blue shaded cells of Table 2), followed by 1st principle (Proactive not Reactive; Preventative not Remedial) with 11 and 6th Principle with 10 maximum values.

Principles with most maximum values had zero minimum values. Least similarity was found between 4th, closely followed by 7th and 3rd Principle. Least similar article was expectedly article §96, which is governing the entry of GDPR into the force and its application. Most similar in the second half was article §87, governing the processing of the national identification number.

Similarities between GDPR's articles 51-96 and 7 principles can be seen in Figure 4. Most similarities are found between GDPR's articles 51-96 and 2nd principle and least between 4th, 7th and 3rd Principle.

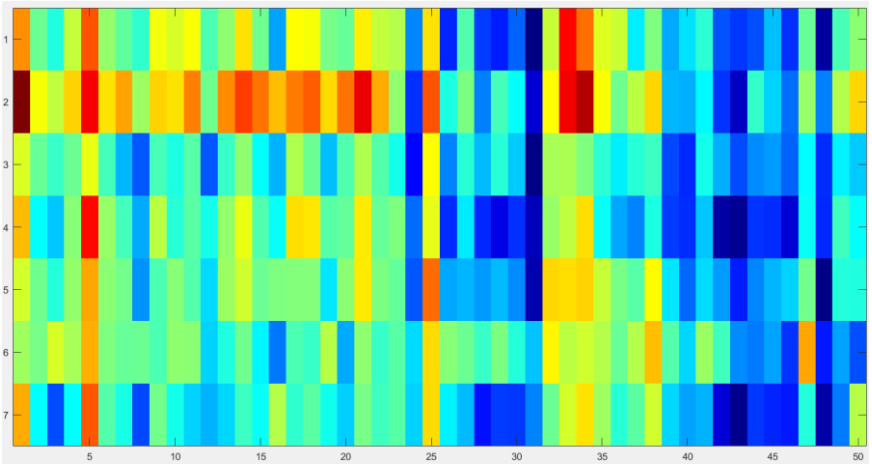

Figure 3: Heatmap of similarities between GDPR's articles 1-50 (horizontal axis) and 7 Privacy by Design principles (vertical axis)

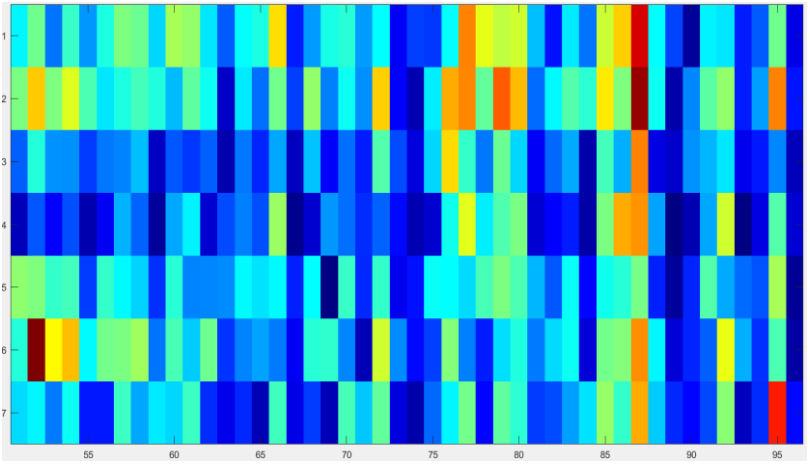

Figure 4: Heatmap of similarities between GDPR's articles 51-96 (horizontal axis) and 7 Privacy by Design principles (vertical axis)

Visual comparison between the first part and the second part of GDPR (Figure 3 and in Figure 4) reveals that, expectedly, the first part is more similar to principles. Values of descriptive statistics (Average, Min and Max values) also reveal that the similarities

between the principles and the text are higher for the first half of GDPR compared to the second half (see Table 3).

Table 3: Comparison of descriptive statistics for 1st and 2nd half of GDPR's similarity measures

	Descriptive	Principle 1	Principle 2	Principle 3	Principle 4	Principle 5	Principle 6	Principle 7
1st half	Average	0,220782	0,277495	0,184459	0,187545	0,210879	0,220821	0,182603
	Max	0,428107	0,491606	0,307547	0,425750	0,376919	0,349675	0,387155
	Min	0,000959	0,033069	0,003259	0,010421	0,003679	0,073135	0,007185
2nd half	Average	0,146231	0,164328	0,097452	0,092645	0,121395	0,138083	0,113533
	Max	0,355063	0,379704	0,290070	0,283847	0,209332	0,389725	0,329203
	Min	0,010111	0,017228	0,016305	0,001147	0,000858	0,013272	0,017552

We have additionally checked whether the differences in means are due to chance alone, hence we used the t-tests. We calculated the $p<0.005$ for each individual principle, meaning we have to reject the null hypothesis of no difference in favor of alternative that there are statistically significant differences in similarity values between part 1 and part 2 of the GDPR.

Finally, for the whole GDPR, we checked whether there are statistically significant differences in similarity values among the principles. We used one-way ANOVA. The calculated F value was $F(6)=9.990$, $p<0.005$, meaning there are differences between the principles.

The location of differences was detected using a post test, namely a Tukey's HSD, which revealed four homogeneous groups (see Figure 5).

Value

Tukey HSD[a]

		Subset for alpha = 0.05			
PrincipleNum	N	1	2	3	4
4	96	,142072534			
3	96	,142768338	,142768338		
7	96	,149506992	,149506992	,149506992	
5	96	,168001484	,168001484	,168001484	
6	96		,181175497	,181175497	
1	96			,185059925	,185059925
2	96				,223269764
Sig.		,428	,053	,096	,056

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 96,000.

Figure 5: Tukey's HSD values among 7 Privacy by Design Principles for the whole GDPR

Interestingly, 2nd Principle was in its own homogeneous group when only the first part of GDPR is considered (ANOVA: $F(6)=7,295$, $p<.000$) – see Figure 6.

Value

Tukey HSD[a]

		Subset for alpha = 0.05	
PrincipleNum	N	1	2
7	50	,182603005	
3	50	,184459649	
4	50	,187545426	
5	50	,210879730	
1	50	,220782596	
6	50	,220821054	
2	50		,277495847
Sig.		,304	1,000

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 50,000.

Figure 6: Tukey's HSD values among 7 Privacy by Design Principles for the first half of GDPR

4. Discussion and conclusion

In this paper, we checked how European regulation on data processing, the famous General Data Protection Regulation (GDPR) directive [14] – the Regulation (E.U.) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data – is addressing the privacy by using the Privacy by Design principles [28].

We used automated text similarities approach and the Universal Sentence Encoder to encode texts of the GDPR's articles and of the 7 Privacy by Design into vectors. Inner product of computed vectors represents the similarity measure among different pairs of texts.

We have found that among all principles, Principle 2 ("*Privacy as the Default*") was most similar, and it was also most strongly similar. Based on average similarity score principles 1 ("*Proactive not Reactive; Preventative not Remedial*"), 6 ("*Visibility and Transparency*") and 5 ("*End-to-End Security – Lifecycle Protection*") followed.

Least similar principles were 4th ("*Full Functionality – Positive-Sum, not Zero-Sum*"), 3rd ("*Privacy Embedded into Design*") and 7th ("*Respect for User Privacy*").

From the order of the principles it is rather surprising that 7th principle was among the least similar to GDPR text, despite the principle urging to build privacy around an individual. Afterall, GDPR is protecting one of the basic human (individual) rights, the right to privacy. On the other hand, automatic semantic analysis has correctly identified Article 5 (describing the principles of GDPR) to be most similar to seven principles of Privacy by Design, followed by Article 1.

Expectedly, the first part of the GDPR was more similar to privacy by design principles than the second part. Interestingly, ANOVA showed that the principle scores were different among each other, and four homogeneous groups formed, with principle 2 and 1 being the most similar to GDPR text, and 4, 3, 7, and 5 least similar. The average difference between the similarity values of the most similar principle #2 (similarity value: 0,223) and the least similar principle #4 (similarity value: 0,142) is more than 60 %.

Our research has shown that automated text similarities approach can discover interesting similarities between legal texts and the underlying principles, not only in general, but in particular for each article–principle pair, or for several portions of texts. The portions of texts where one or several principles prevail can easily be uncovered using visual representation of similarity scores and differences checked using traditional statistical methods.

Acknowledgement

The author acknowledges the financial support from the Slovenian Research Agency (research core funding No. P2-0057, project funding No. V5-1725) and from University of Maribor (www.um.si, core funding).

References

[1] Jowett B. Complete works of Aristotle. In: Barnes J, editor. Princeton, NJ: Princeton University Press; 1995.
[2] Roy J. 'Polis' and 'Oikos' in Classical Athens. Greece & Rome. 1999;46(1):1-18.
[3] Shields JM. A Sacrifice to Athena: Oikos and Polis in Sophoclean Drama. Lewisburg, PA: Bucknell University, Department of Religion; 1991; Available from: http://www.facstaff.bucknell.edu/jms089/Z-Unpublished%20Work/Athena.pdf. (Archived by WebCite® at http://www.webcitation.org/6Axq3vRN7).
[4] Post S.G. Encyclopedia of Bioethics. New York, U.S.A.: Macmillan Reference; 2004. ISBN: 9780028657783.
[5] Brumen B, Heričko M, Sevčnikar A, Završnik J, Hölbl M. Outsourcing medical data analyses: can technology overcome legal, privacy, and confidentiality issues? J Med Internet Res. 2013 December 16, 2013;15(12):e283. PMID: 24342053. doi: 10.2196/jmir.2471.
[6] Rooney B. Facebook Understands Europe's Privacy Fears Says Sandberg. The Wall Street Journal, TechEurope 2013-04-19.
[7] Segall L. Facebook was 'the first innovator in privacy,' C.O.O. says. CNN Money. 2011-12-01.
[8] Cadwalladr C, Graham-Harrison E. Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. The guardian. 2018;17:22.
[9] Isaak J, Hanna MJ. User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection. Computer. 2018;51(8):56-9. doi: 10.1109/MC.2018.3191268.
[10] Bond RM, Fariss CJ, Jones JJ, Kramer A.D.I., Marlow C, Settle JE, et al. A 61-million-person experiment in social influence and political mobilization. Nature. 2012 2012/09/01;489(7415):295-8. doi: 10.1038/nature11421.
[11] McKenzie S. Facebook's Mark Zuckerberg says sorry in full-page newspaper ads. CNN. 2018 March 25, 2018;Sect. Europe.
[12] Allen AL. Privacy Law and Society. 1st ed: Thomson West; 2007 August 31, 2007. ISBN: 0314163581.
[13] Allen AL. Privacy and Medicine. In: Zalta EN, editor. The Stanford Encyclopedia of Philosophy (Spring 2011 Edition). Stanford, CA, U.S.A.: Stanford University; 2011.
[14] EU. Regulation (E.U.) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data. Official Journal of the European Union. 2016 4.5.2016;L:2016:119.
[15] Solove D.J. A Taxonomy of Privacy. U Pa L Rev. 2006;154(3):477-564.
[16] Warren SD, Brandeis LD. The Right to Privacy. Harv Law Rev. 1890;4(5):193-220.
[17] DeCew JW. Privacy. In: Zalta EN, editor. The Stanford Encyclopedia of Philosophy. Stanford, CA, U.S.A.: Stanford University; 2008.
[18] Prosser WL. Privacy. Calif Law Rev. 1960;48(3):383-423.
[19] Nagel T. Concealment and exposure: and other essays. New York, U.S.A.: Oxford University Press; 2002. ISBN: 019515293X.
[20] Moore AD. Privacy: its meaning and value. American Philosophical Quarterly. 2003;40(3):215-27.
[21] Gavison R. Privacy and the Limits of Law. Yale L J. 1980;89(3):421-71.
[22] Allen AL. Uneasy access: Privacy for women in a free society. Totowa, NJ, U.S.A.: Rowman & Littlefield Pub Inc; 1988. ISBN: 0847673286.
[23] Bok S. Secrets: On the ethics of concealment and revelation. New York, U.S.A.: Vintage; 1989. ISBN: 0679724737.
[24] DeCew JW. The priority of privacy for medical information. Soc Philos Policy. 2000;17(2):213-34.
[25] Allen AL. Genetic Privacy: Emerging Concepts and Values. In: Rothstein MA, editor. Genetic secrets: Protecting privacy and confidentiality in the genetic era. New Haven: Yale University Press; 1997.
[26] Kenny DJ. Confidentiality: the confusion continues. J Med Ethics. 1982 March 1, 1982;8(1):9-11. doi: 10.1136/jme.8.1.9. PMID: 7069738.
[27] Sharpe VA. Privacy and Security for Electronic Health Records. Hastings Cent Rep. 2005;35(6):c3. doi: 10.1353/hcr.2005.0115. PMID: 16396204.
[28] Cavoukian A. Privacy by design: The 7 foundational principles. Information and privacy commissioner of Ontario, Canada. 2009.
[29] Cavoukian A. Big Data & Privacy Together – It Is Achievable. Office of the Privacy Commissioner (Ontario): Ontario, Canada. Ontario, Canada: Office of the Privacy Commissioner (Ontario); 2013; Available from: http://www.privacybydesign.ca/index.php/big-data-privacy-together-is-achievable/. (Archived by WebCite® at http://www.webcitation.org/6GpVPLaXK).
[30] Cavoukian A. 7 Foundational Principles of Privacy By Design. Ontario, Canada.: Office of the Privacy Commissioner (Ontario); 2013; Available from: http://www.privacybydesign.ca/index.php/about-pbd/7-foundational-principles/. (Archived by WebCite® at http://www.webcitation.org/6GqwiKldy).

[31] Cavoukian A, Chanliau M. Privacy and Security by Design: A Convergence of Paradigms. Ontario, Canada: Office of the Privacy Commissioner (Ontario), 2013.
[32] Cavoukian A. Privacy by design [leading edge]. IEEE Technology and Society Magazine. 2012;31(4):18-9.
[33] Cuijpers C, Purtova N, Kosta E. Data protection reform and the Internet: the draft Data Protection Regulation. Research Handbook on E.U. Internet Law: Edward Elgar Publishing; 2014.
[34] van Dijk N, Gellert R, Rommetveit K. A risk to a right? Beyond data protection risk assessments. Computer Law & Security Review. 2016;32(2):286-306.
[35] EDPS. Opinion 5/2018: Preliminary Opinion on privacy by design. Brussels, Belgium: European Data Protection Supervisor, 2018.
[36] Kuner C, Bygrave LA, Docksey C, editors. The E.U. General Data Protection Regulation (GDPR). A Commentary. Oxford, United Kingdom: Oxford University Press; 2020.
[37] Cer D, Yang Y, Kong S-y, Hua N, Limtiaco N, John RS, et al., editors. Universal sentence encoder for English. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations; 2018.
[38] Sieg A. Text Similarities : Estimate the degree of similarity between two texts. 2018; Available from: https://medium.com/@adriensieg/text-similarities-da019229c894. Archived by archive.ph at http://archive.ph/VDtpY. Last accessed Feb/2020.
[39] Cer D, Diab M, Agirre E, Lopez-Gazpio I, Specia L. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In: Bethard S, Carpuat M, Apidianaki M, Mohammad SM, Cer D, Jurgens D, editors. Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2017); August 2017; Vancouver, Canada: Association for Computational Linguistics; 2017. doi: 10.18653/v1/S17-2001.