Automated GDPR-Compliance in Requirements Engineering

Abdel-Jaouad Aberkane[0000−0002−4557−0715]

Department of Business Informatics and Operations Management, Faculty of Economics and Business Administration, Ghent University, Ghent, Belgium
abdeljaouad.aberkane@ugent.be

Abstract. In the last lustrum, the EU General Data Protection Regulation (GDPR) profoundly impacted data processing organizations as compliance with this Regulation became obligatory. Due to resource poverty, complying with the GDPR can be a challenge for small and medium-sized enterprises. In this research, we consider GDPR-compliance as a high-level goal in software development that should be addressed at the outset of software development, that is, during requirements engineering (RE). Moreover, we argue that natural language processing (NLP) can be utilized to automate this process. Therefore, this Ph.D. research aims to address the challenge organizations face by developing an NLP-based automated approach towards GDPR-compliance in RE. In particular, we aim to develop an approach to assess whether a set of system requirements complies with the GDPR to achieve data protection by design and by default, thus providing organizations with an efficient and effective solution to ensure GDPR-compliance. This paper presents our research questions and their relevance, the adopted research method, preliminary results, and the current state of our research.

Keywords: General Data Protection Regulation · Machine Learning · Natural Language Processing · Requirements Engineering

1 Introduction

Provisionally agreed upon in 2015, the General Data Protection Regulation (GDPR) came into effect in May 2018 to ensure and safeguard data subjects' rights. The GDPR applies to any personal data processing of European data subjects—regardless of whether the data processing itself takes place in the European Union—and ensures that individuals are not deprived of their data protection rights. Furthermore, the GDPR's transnational application results in far-reaching implications on the digital market of the world [2].

To achieve GDPR compliance, organizations should take technical and organizational measures necessary to ensure data protection by design and by default to achieve GDPR-compliance [16]. Since the conditions for data processing are

fundamentally being set by the soft- and hardware used for the task [19], we consider GDPR-compliance to be a high-level goal in software development. In this Ph.D. research, we aim to address GDPR-compliance at the outset of software development, meaning requirements engineering, to achieve data protection by design and by default. Requirements engineering consists of the elicitation, evaluation, specification, analysis, and evolution of the objectives, functionalities, qualities, and constraints to be achieved by a software-intensive system [18]. In this research, these constraints are set by the GDPR.

Complying with the GDPR demands significant efforts, which can be especially troublesome for small and medium-sized enterprises (SMEs) due to lack of resources to take the necessary technical, and organizational measures [8, 17]. Moreover, the need to find efficient and effective solutions to ensure GDPR-compliance [8] is emphasized by the fact that the vast majority of businesses (99%) in the European Union are SMEs [7].

Therefore, we propose using natural language processing (NLP) techniques—a popular approach amongst requirements engineers [4]—as a means of automating GDPR-compliance in organizations. Requirements engineering is especially conducive for the use of NLP since requirements are often expressed in natural language [11].

The main focus of this Ph.D. research is to prevent organizations from compromising the privacy of data subjects and ensuring data protection by design and by default, by addressing GDPR-compliance at the intersection of RE and NLP. The remainder of this paper is structured as follows. Section 2 presents our research questions. Section 3 describes the adopted research method. The preliminary results are presented in Section 4. Section 5 presents the current stage of our research. Related work is presented in Section 6. Finally, Section 7 concludes this paper and elaborates on our plans for future work.

2 Research Questions

The goal of this Ph.D. research is to achieve NLP-based automated GDPR-compliance in RE. In doing so, we aim to build a bridge between GDPR, NLP, and RE to provide organizations with insights towards GDPR-compliance. Furthermore, we intend to develop a machine learning approach based on NLP to facilitate GDPR-compliance and prevent organizations from compromising EU citizens' privacy, thus ensuring data protection. To achieve this, we adopt an incremental approach which is reflected in our research questions (RQs) and the corresponding sub-research questions (SRQs):

RQ1 "*What is the current state of research as to NLP-based automated GDPR-compliance in RE?*" This RQ aims to map the existing literature on the intersection of GDPR, NLP and RE relevant to achieving automated GDPR-compliance. However, since preliminary research yielded no relevant studies, it was decided to formulate SRQs to search all possible combinations of the three domains (i.e., GDPR, NLP, and RE):

SRQ1 "*What NLP approaches are useful for RE and for which RE activity?*" Literature shows that several NLP approaches are available for the different activities in RE. However, no systematic literature mapping exists where these different approaches are mapped with the tasks they can be used for while upholding the higher goal of automating GDPR-compliance in RE.
SRQ2 "*Which NLP approaches are available for achieving GDPR-compliance in organizations?*" In addition to mapping the NLP approaches explicitly developed for the RE domain, it is also interesting to explore NLP approaches outside the RE paradigm that focus on achieving GDPR-compliance in organizations. This insight will allow us to investigate the possibility of learning from these approaches and transfer this learning to the RE phase in software development.
SRQ3 "*Which state-of-the-art RE solutions are available for achieving GDPR-compliance?*" This question endeavors to identify state-of-the-art RE solution types that maintain and possibly facilitate the GDPR. The answer to this question will provide us with a topical mapping of the RE developments in achieving GDPR-compliance.

RQ2 "*How to automatically assess organizations' GDPR-compliance based on their privacy policy?*" This research question aims to assess organizations' GDPR-compliance through an NLP-based machine learning approach. Organizations use privacy policies to convey the taken GDPR-measures to data subjects. We will particularly focus on SMEs and provide them with an efficient and effective approach to identify possible GDPR non-compliance.
RQ3 "*How do organizations differ in their approach to GDPR-compliance?*" Building on the previous RQ, this question aims to identify the influence of organizational factors such as size and industry on GDPR-compliance. The construed machine learning approach of RQ2 will be used to classify a data set comprising organizations and their respective privacy policies on their GDPR compliance. The resulting classification will be combined with organizational meta-data of the corresponding organizations, after which possible relationships between GDPR-compliance and organizational factors will be identified through machine learning. The outcome may validate the need for a tailored approach towards GDPR-compliance for organizations, addressed in RQ4.
RQ4 "*How to automatically identify the GDPR-compliance of a set system requirements?*" This question is built on the outcomes of RQ2 and RQ3 and aims to devise an automated approach based on NLP and machine learning, tailored to the outcome of RQ3, that can be used to assess, by using NLP-based machine learning, the GDPR-compliance of a set system requirements at the outset of software development and thus achieving data protection by design and by default.

Summarized, we adopt an incremental approach towards our goal of NLP-based automated GDPR-compliance in RE. First, we intend to explore the literature on the intersection of GDPR, NLP, and RE. Next, we aim to develop an NLP-based machine learning approach to assess the GDPR-compliance of organizations' privacy policies. Furthermore, we aim to identify the organiza-

tional factors that influence this GDPR-compliance. Based on the identified factors, we intend to devise an automated approach based on machine learning that focuses on achieving GDPR-compliance by design and by default, by automatically assessing system requirements at the beginning of the software development process on their GDPR-compliance. This approach will enable requirements engineering professionals to identify requirements automatically that may lead to non-compliance, consequently instilling a proactive stance towards GDPR-compliance. The envisaged outcomes and solutions are intended to facilitate GDPR-compliance in organizations by providing the corresponding professionals with the means to achieve this compliance.

3 Research Method

This research adopts the *design science research methodology*. Design science research is a research paradigm that focuses on answering questions related to human problems by creating new and innovating artifacts [9]. This research paradigm is built upon the fundamental principle that knowledge and understanding of a design problem and its solutions are acquired in the building and application of an artifact. In examining how artifacts can be developed, design science research can be seen as an embodiment of three closely related cycles of activities: the relevance cycle, the rigor cycle, and the design cycle [10]. The relevance cycle links the contextual environment with the research and introduces the research artifacts into environmental field testing. The rigor cycle bridges the knowledge base with the research and adds incrementally the new knowledge created by the research to the knowledge base. Lastly, the design cycle iterates between the core activities of building and evaluating the design artifacts and processes of the research.

In this Ph.D. research, the need to find efficient and effective solutions to ensure GDPR-compliance in organizations forms the problem space of interest. We intend to address this by developing an NLP-based machine learning approach to assist organizations in meeting GDPR-compliance in RE. However, first, we aim to validate this need and understand the problem space by developing an artifact that assesses GDPR-compliance of organizations based on the measures adopted as communicated through their privacy policies, followed by identifying the organizational factors that contribute to non-compliance—thus iterating between the relevance cycle and the design cycle while contributing to the knowledge base. Based on the outcome, we intend to develop a machine learning approach—in the design cycle—to achieve GDPR-compliance by design and by default, by automatically assessing system requirements in the RE phase of software development on their GDPR-compliance. Finally, we aim to evaluate the performance of this machine learning approach using expert-labeled data.

4 Preliminary Results

This Ph.D. research started with conducting a systematic mapping study [1], following the guidelines in [12], to address RQ1 and its sub-research questions and identify the current research state of NLP-based automated GDPR-compliance in RE. The mapping study resulted in 448 relevant studies. The majority of these studies—420 studies—were identified as relevant to SRQ1, which centers around NLP and RE. SRQ2, focusing on GDPR and NLP, yielded nine studies. SRQ3, which centers around GDPR and RE, resulted in 20 studies. One of the retrieved studies fell within all pairwise research field combinations (i.e., NLP & RE, NLP & GDPR, and GDPR & RE) [15]. The referenced study presents a recommender-based privacy requirements elicitation approach (EPICUREAN) which uses NLP and machine learning techniques in the RE activity of domain understanding & elicitation to determine and recommend appropriate privacy settings to the user and, as a result of this, simplifying privacy settings according to the GDPR.

Regarding SRQ1, 420 studies were identified, mentioning 199 different NLP approaches useful for RE. The most common NLP approaches were *part-of-speech tagging*, *pre-processing*, and *text-classification* with 165, 117, and 104 occurrences, respectively. SRQ2, aimed to identify studies that discuss NLP approaches used to achieve GDPR-compliance in organizations, yielded nine studies in which text-classification and pre-processing were the most frequently occurring NLP approaches with 4 and 2 occurrences, respectively. Furthermore, the retrieved studies related to SRQ2 addressed the GDPR concept of *anonymization* four times, *privacy* was addressed twice, *consent* and *lawfulness, fairness and transparency* occurred each once. Lastly, regarding SRQ3, 20 studies were identified that discussed RE solutions for achieving GDPR-compliance in organizations. The most frequent occurring solutions types were *Approach* and *Method*, proposed as a solution by four studies each.

Next to the key findings that addressed our research questions, we discovered several interesting phenomena. First, the mapping revealed a tendency among researchers from the RE community to concentrate mainly on the RE activities of domain understanding & elicitation and specification & documentation. Concurrently, the activities of evaluation & negotiation, and quality assurance receive less attention. This development was true for research related to both SRQ1 and SRQ2. For studies related to SRQ1, this may imply that researchers overlooked the possibility of using NLP for the activities of evaluation & negotiation, and quality assurance. Another possibility is that RE tasks in the latter are less suitable for automating with NLP. Second, it could be foreseen from the line of questioning in our research questions that the resulting studies will gravitate towards design science research. Therefore, it is not unexpected that the vast majority of studies focused on proposing a solution and acted within this research paradigm. However, this is done at the expense of other research types, to the detriment of the diversity of RE as a research domain.

Despite identifying only one study on the intersection of GDPR, NLP, and RE, we have—during the mapping process—identified possibilities for bridging

these research fields to achieve GDPR-compliance through data protection by design and by default in RE. In particular, we focus on literature related to SRQ2 and SRQ3 because they address the GDPR explicitly. Scrutinizing the solutions proposed by literature related to NLP and GDPR (SRQ2) gives rise to the opportunity of introducing these approaches to classification problems in the RE domain, for instance, using NLP-based machine learning solutions that identify GDPR-compliance on requirement documents. Moreover, examining the studies related to SRQ3 presents possibilities of using NLP techniques to assist with manual and potentially repetitive tasks on the junction of GDPR and RE.

5 Current Stage

At this stage, we are working towards an approach to automatically assess organizations' GDPR-compliance based on their privacy policy (RQ2). As mentioned in Section 2, this question aims to assess organizations' GDPR-compliance by developing an NLP-based machine learning approach that can be used to assess privacy policies' GDPR-coverage based on a set of predefined assessment criteria. The results of such a classification may imply that a particular organization fails to comply with the GDPR-measures. Specifically, we will focus on SMEs—due to the threat of resource poverty—and provide them with an efficient and effective approach to identify possible GDPR non-compliance. Currently, we are concerned with collecting and labeling data that can be used to train our machine learning model. After this process, we aim to build the classification model and evaluate its usefulness. Consequently, we will extend this research to RQ3, which focuses on identifying how organizations differ in their approach to GDPR-compliance. In detail, we aim to use machine learning techniques to identify factors (e.g., organization size) that influence GDPR non-compliance.

6 Related Work

As manifested through our mapping study, research related to automating GDPR-compliance is still in its infancy. Due to the relative novelty of the GDPR (i.e., it is enforced since 2018), not much research was conducted on the intersection of GDPR, NLP, and RE. In fact, our mapping study (See Section 4) identified only one study on the convergence of GDPR, NLP, and RE [15]. The authors present a recommender-based privacy requirements elicitation approach that uses natural language processing techniques to determine and provide—in line with the GDPR—users with privacy setting suggestions tailored to their needs. Our research does not focus on eliciting requirements; instead, we aim to automatically assess system requirements on their GDPR-compliance using NLP-based machine learning. The following paragraph will shed light on some contiguous research that centers on using NLP towards GDPR-compliance.

In [14], a new model auditing technique is presented to help enforce data protection regulations (e.g., GDPR) by aiding users in detecting whether their data was used to train a machine learning model. In the same stream, [3] proposes

an automated privacy policy extraction system that considers users' privacy concerns while providing the user with GDPR items to assist in privacy-aware decision making. Reference [5] proposes a supervised machine learning system that identifies personal information in large data sets to ease processing such data in an organization according to the GDPR. Along the same lines, [6] presents a system based on NLP and machine learning capable of correctly identifying sensitive data, assisting companies in complying with standards such as the GDPR. Moreover, [13] presents an approach that implements NLP to detect personally identifiable information in contracts.

However, the studies above are located on the crossing of GDPR and NLP, whereas our research aims to use NLP to automate GDPR-compliance in RE to achieve data protection by design and default, thus combining the fields of GDPR, NLP, *and* RE.

7 Conclusion and Future Work

This paper presents our research approach to achieve data protection by design and by default through developing an NLP-based machine learning approach to assess system requirements on their GDPR-compliance. In preparation for developing this approach, we have conducted a systematic mapping study to outline the current research state on the junction of GDPR, NLP, and RE. The mapping study shows potential for our intersectional research. Currently, we are devising a machine learning approach to assess organizations' privacy policies on GDPR-compliance to validate the need for data protection by design and by default as intended by our primary objective. Thereafter, we aim to identify organizational factors that influence GDPR-compliance. Finally, we plan to develop an NLP-based machine learning approach that focuses on achieving GDPR-compliance by automatically assessing the compliance of system requirements at the beginning of the software development process. In summary, this Ph.D. research contributes to a novel research stream on the intersection of GDPR, NLP, and RE by developing an NLP-based automated approach towards GDPR-compliance in RE and towards ensuring data protection by design and by default.

Acknowledgments

This Ph.D. research is supervised by Prof. Dr. Geert Poels and Dr. Seppe vanden Broucke.

References

1. Aberkane, A.J., Poels, G., Broucke, S.V.: Exploring automated gdpr-compliance in requirements engineering: A systematic mapping study. IEEE Access **9**, 66542–66559 (2021). https://doi.org/10.1109/ACCESS.2021.3076921
2. Albrecht, J.P.: How the gdpr will change the world. Eur. Data Protection Law Rev. **2**(3), 287 – 289 (2016). https://doi.org/10.21552/edpl/2016/3/4

3. Chang, C., Li, H., Zhang, Y., Du, S., Cao, H., Zhu, H.: Automated and personalized privacy policy extraction under gdpr consideration. In: International Conference on Wireless Algorithms, Systems, and Applications. pp. 43–54. Springer, Cham (Jun 2019)
4. Dalpiaz, F., Ferrari, A., Franch, X., Palomares, C.: Natural language processing for requirements engineering: the best is yet to come. IEEE Softw. **35**(5), 115–119 (Sep 2018). https://doi.org/10.1109/MS.2018.3571242
5. Di Cerbo, F., Trabelsi, S.: Towards personal data identification and anonymization using machine learning techniques. In: European Conference on Advances in Databases and Information Systems, vol. 909, pp. 118–126. Springer, Cham (Aug 2018)
6. Dias, M., Ferreira, J.C., Maia, R., Santos, P., Ribeiro, R.: Privacy in text documents. In: Proceedings of the 33rd International Business Information Management Association Conference. pp. 2551–2560 (Apr 2019)
7. European Commission: Internal market, industry, entrepreneurship and smes, https://ec.europa.eu/growth/smes/sme-definition_en, accessed Apr. 9, 2021
8. Freitas, M.d.C., Mira da Silva, M.: Gdpr compliance in smes: There is much to be done. J. Inf. Syst. Eng. & Manage. **3**(4), 30 (2018). https://doi.org/10.20897/jisem/3941
9. Hevner, A., Chatterjee, S.: Design science research in information systems. In: Design research in information systems, pp. 16–19. Integrated Series in Information Systems, Springer, Boston, MA, USA (2010). https://doi.org/10.1007/978-1-4419-5653-8
10. Hevner, A.R.: A three cycle view of design science research. Scandinavian J. Inf. Syst. **19**(2), 4 (2007)
11. Kassab, M., Neill, C., Laplante, P.: State of practice in requirements engineering: contemporary data. Innov. Syst. Softw. Eng. **10**(4), 235–241 (Apr 2014). https://doi.org/10.1007/s11334-014-0232-4
12. Keele, S.: Guidelines for performing systematic literature reviews in software engineering. Tech. Report 2.3, Dept. Comput. Sci., Univ. Durham, Durham, U.K. (Jul 2007)
13. Silva, P., Gonçalves, C., Godinho, C., Antunes, N., Curado, M.: Using natural language processing to detect privacy violations in online contracts. In: Proceedings of the 35th Annual ACM Symposium on Applied Computing. pp. 1305–1307. ACM, New York, NY, United States (Mar 2020)
14. Song, C., Shmatikov, V.: Auditing data provenance in text-generation models. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 196–206. ACM, New York, NY, USA (Jul 2019)
15. Stach, C., Steimle, F.: Recommender-based privacy requirements elicitation - EPICUREAN. In: Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing. pp. 1500–1507. ACM, New York, NY, USA (Apr 2019)
16. The European Parliament and the Council of European Union: REGULATION (EU) 2016/679. Official Journal of the European Union (2016). https://doi.org/http://data.europa.eu/eli/reg/2016/679/2016-05-04
17. Tikkinen-Piri, C., Rohunen, A., Markkula, J.: Eu general data protection regulation: Changes and implications for personal data collecting companies. Comput. Law & Sec. Review **34**(1), 134–153 (2018). https://doi.org/10.1016/j.clsr.2017.05.015
18. Van Lamsweerde, A.: Requirements engineering: From system goals to UML models to software, vol. 10, pp. 30–34. John Wiley & Sons, Chichester, UK (2009)

19. Voigt, P., von dem Bussche, A.: The EU General Data Protection Regulation (GDPR), vol. 1, p. 62. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-57959-7