A First Look at the General Data Protection Regulation (GDPR) in Open-Source Software

Lucas Franke
Virginia Tech
Blacksburg, VA, USA
lfranke@vt.edu

Huayu Liang
Virginia Tech
Blacksburg, VA, USA
huayu98@vt.edu

Aaron Brantly
Virginia Tech
Blacksburg, VA, USA
abrantly@vt.edu

James C. Davis
Purdue University
W Lafayette, IN, USA
davisjam@purdue.edu

Chris Brown
Virginia Tech
Blacksburg, VA, USA
dcbrown@vt.edu

ABSTRACT

This poster describes work on the General Data Protection Regulation (GDPR) in open-source software. Although open-source software is commonly integrated into regulated software, and thus must be engineered or adapted for compliance, we do not know how such laws impact open-source software development.

We surveyed open-source developers (N=47) to understand their experiences and perceptions of GDPR. We learned many engineering challenges, primarily regarding the management of users' data and assessments of compliance. We call for improved policy-related resources, especially tools to support data privacy regulation implementation and compliance in open-source software.

ACM Reference Format:
Lucas Franke, Huayu Liang, Aaron Brantly, James C. Davis, and Chris Brown. 2024. A First Look at the General Data Protection Regulation (GDPR) in Open-Source Software. In *Proceedings of International Conference on Software Engineering — Invited Posters (ICSE'24-Poster).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 PROBLEM STATEMENT

Software products collect user data to enhance user experiences through personalized, machine learning-enabled [7] application behaviors [5]. This may both benefit users and threaten their well-being, *e.g.,* swaying elections in the USA [17]. To protect their citizens, over 100 governments worldwide are developing *data privacy laws and regulations* to constrain how citizens' personal data is collected, processed, stored, and saved [15]. A landmark example is the EU's General Data Protection Regulation (GDPR), which grants rights to citizens for the handling of their data [1].

Data privacy regulations create challenging software requirements because they entail both technical and legal expertise. Software developers may have limited legal knowledge [10, 16] and receive minimal training [3, 8]. However, there has been limited study of how such laws affect the software development process. The few existing studies have been of commercial software development [2, 4]; we lack knowledge of the effects of GDPR and other regulations on open-source software (OSS) development.

You may be surprised that regulatory compliance is a factor in OSS development, as most OSS licenses disclaim legal responsibility. However, users and developers of OSS may still desire regulatory compliance. We note three examples. (1) A majority of OSS is developed for commercial use [6] and may require standards or regulatory compliance. (2) Companies who integrate OSS components in their software supply chains [11] may request the addition of compliance requirements, and the developers may service these requests. (3) Companies may find the (free) price of OSS compelling and undertake their own compliance analysis [13]. Standards such as IEC 61508–Part 3 include provisions for doing so [9]. In a 2023 survey of ~1700 codebases across 17 industries, Synopsys found OSS in 96% of the codebases and reported an average contribution of 75% of the code in each codebase [14].

We therefore explore GDPR compliance in open-source software.

2 APPROACH

GDPR compliance is an unstudied topic in OSS. We therefore adopt an exploratory methodology, focused on qualitative data, to provide a characterization and identify phenomena of interest. We followed a four-step approach aligned with the framework analysis methodology [12]. Since we studied human subjects, our Institutional Review Board (IRB) provided oversight.

Step 1: Pilot Study and Data Familiarization: To formulate an initial thematic framework for our qualitative analysis, we conducted semi-structured pilot interviews with three developers.

Two researchers coded the interview transcripts to extract themes (examples in Table 1). Participants highlighted the challenges with implementing GDPR requirements in open-source software, including (1) legal consultation, and (2) compliance assessment.

Step 2: Survey Design: Our pilot study findings informed our survey design. As summarized in Table 1, we asked short-answer questions about GDPR in OSS. We asked respondents about the perceived impact of the GDPR on data privacy, difficult concepts to implement, and GDPR compliance assessment.

Step 3: Participant Recruitment: We wanted respondents with experience implementing GDPR in OSS. We searched GitHub for repositories with GDPR-related pull requests (searching for "GDPR" typically yields English-language PRs related to GDPR compliance). We emailed developers who authored or commented on these PRs. From 98 emails we received 5 responses (5% response rate). We then made broader calls for participation on Twitter and Reddit.

We received a total of 47 responses that self-reported experience implementing GDPR compliance. Participants have a median of ~5 years of OSS development experience and 6 years of industry experience. Participants contributed to OSS projects such as Ansible, Django, Kubernetes, PostGreSQL, and Microsoft Cognitive Toolkit.

Step 4: Data Analysis: To analyze our survey results, we used an open coding approach. Two analysts independently performed a

ICSE'24-Poster, Apr.'24, Lisbon, Portugal

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Table 1: Sample questions from pilot study and survey. The final column notes the inter-rater agreement score for these themes (Cohen's κ).

Interview Question	Example Codes	Survey Question	Example Codes	κ
What impact, if any, do you believe the GDPR has had on data security and privacy?	data privacy, rights to users, data collection	What impact, if any, do you believe the GDPR and similar data privacy regulations have had on data security and privacy?	data privacy, data processing, data collection, insufficient information, data breach, fines	0.736
What GDPR concepts do you find the most difficult or frustrating to implement?	None, data minimization, embedded content	What GDPR concepts do you find the most difficult or frustrating to implement?	privacy by design, data minimization, cost, data processing, user experience	
Have you sought legal consultation on GDPR-related issues, and if so, how did that affect your development process?	Yes/No; no effect, negative effect (time)	Have you sought legal consultation on GDPR-related issues, and if so, how did that affect your development process?	Yes/No; no effect; positive; negative (cost, time, data storage, data processing,...)	0.514

manual inspection of responses, categorizing responses based on the themes derived from our pilot study. If a new theme arose, the analysts discussed before adding it. We used Cohen's kappa (κ) to measure inter-rater agreement (Table 1).

3 RESULTS SO FAR

Positive Views: Six participants had positive perceptions of the GDPR. For example, they said "*the risk of incurring and paying out hefty fines has made companies take privacy and security more proactively*" (P30), that GDPR brings "*awareness to the importance about privacy*" (P45), and that "*data integrity is ensured*" (P47). These responses reflect the intentions of the GDPR.

Negative Views: 16 participants reported negative views. These responses primarily focused on three issues: cost, organizations, and enforcement. For costs, respondents noted that implementing GDPR requirements is expensive and burdensome. Participants said that compliance is "*costly for many companies*" (P16) is "*too expensive*" (P24), and "*the cost of protection should not go over the cost of consequence of data breach...GDPR [isn't] worth the time*" (P46). For organizations, participants reported a negative impact of the GDPR on companies and organizations. They mentioned that GDPR compliance "*weakens small and medium-sized enterprises*" (P15), "*threatens innovation*" (P18), and "*fails to meaningfully integrate the role of privacy-enhancing innovation and consumer education in data protection*" (P23). P46 added that the GDPR is "*a lot of headache...jobs for lawyers at the expense of people who are trying to solve real problems*". For enforcement, one subject said "*there is a large gap in GDPR enforcement among member states* (P17) and another had observed changes over time — "*the trend...is an increase in the number of times and the amount of fines*" (P18).

Engineering Challenges: We summarize three categories here.

(1) Software Design: P21 felt that GDPR compliance reduced the quality of their application's design: "*the principle of minimum scope was not observed*". One participant said GDPR work showed "*things we had not considered before*", *e.g.,* "*logging functionality*" and "*access restrictions*". P17 observed difficulties with new technologies: "*GDPR's requirements are essentially incompatible with big data, artificial intelligence, blockchain, and machine learning*".

(2) Legal Compliance: 11 respondents reported consulting with legal teams for GDPR compliance. Most of them (7) lamented this need, stating "*it slows things down as code has to be reviewed and objectives revised*" and "*it impacted our approach to the SDLC*" (P1), "*it's a bit of a headache*" (P24), and the development process was affected "*if the development is production software*" (P27) already deployed to users. P47 stated "*open-source projects can't afford even to sustain maintainers, not even speaking about legal team*".

(3) Validation Lacking legal counsel, many respondents felt responsible for evaluating "*legality*" (P18) and "*integrity and confidentiality*" (P23) of user data processing and storage. P24 responded developers have to "*consider whether you really need all the data you collect*" and P38 advised to "*get your consent in order*". P18 added there is "*really no good way*" to evaluate compliance.

REFERENCES

[1] 2016. Regulation (EU) 2016/679 of the European Parliament. Official Journal of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679

[2] Abdulrahman Alhazmi and Nalin Asanka Arachchilage. 2021. I'm all ears! listening to software developers on putting GDPR principles into software development practice. *Personal and Ubiquitous Computing* 25 (2021), 879–892.

[3] Keri Allan. 2007. Reskilling for compliance. *Information Professional* 4, 1 (2007).

[4] Kathrin Bednar, Sarah Spiekermann, and Marc Langheinrich. 2019. Engineering Privacy by Design: Are engineers ready to live up to the challenge? *The Information Society* 35, 3 (2019), 122–142.

[5] Randolph E Bucklin and Catarina Sismeiro. 2009. Click here for Internet insight: Advances in clickstream data analysis in marketing. *Journal of Interactive marketing* 23, 1 (2009), 35–48.

[6] GitHub. 2022. Octoverse 2022: The state of open source software. https://octoverse.github.com

[7] Paul Hitlin, Rainie Lee, and Kenneth Olmstead. 2019. Facebook Algorithms and Personal Data. Pew Research Center.

[8] Sebastian Holst. 2017. GDPR liability: software development and the new law. *LinkedIn* (2017). https://www.linkedin.com/pulse/gdpr-liability-software-development-new-law-sebastian-holst/

[9] International Electrotechnical Commission. 2010. Functional safety of electrical/electronic/programmable electronic safety-related systems - Part 3: Software requirements. https://webstore.iec.ch/publication/9277

[10] Rene Moquin and Robin L Wakefield. 2016. The roles of awareness, sanctions, and ethics in software compliance. *Journal of Computer Information Systems* 56, 3 (2016), 261–270.

[11] Chinenye Okafor, Taylor R Schorlemmer, Santiago Torres-Arias, and James C Davis. 2022. Sok: Analysis of software supply chain security by establishing secure design properties. In *Proceedings of the 2022 ACM Workshop on Software Supply Chain Offensive Research and Ecosystem Defenses*. 15–24.

[12] Jane Ritchie and Liz Spencer. 2002. Qualitative data analysis for applied policy research. In *Analyzing qualitative data*. Routledge, 173–194.

[13] David Stokes. 2012. 21 - Validation and regulatory compliance of free/open source software. In *Open Source Software in Life Science Research*, Lee Harland and Mark Forster (Eds.). Woodhead Publishing, 481–504.

[14] Synopsys. 2023. Open Source Security and Risk Analysis Report. https://www.pwc.com/us/en/services/consulting/library/gdpr-readiness.html

[15] UNCTAD. 2021. Data Protection and Privacy Legislation Worldwide. *United Nations Conference on Trade and Development* (2021).

[16] Denis Verdon. 2006. Security policies and the software developer. *IEEE Security & Privacy* 4, 4 (2006), 42–49.

[17] Christopher Wylie. 2019. How I Helped Hack Democracy. New York Magazine.