Selected Papers of #AoIR2022:
The 23rd Annual Conference of the
Association of Internet Researchers
Dublin, Ireland / 2-5 Nov 2022

VOICES FROM THE MARGINS: PRIVACY DISCOURSE IN GITHUB README FILES

Keren Levy Eshkol
Department of Communication, University of Haifa

Rivka Ribak
Department of Communication, University of Haifa

As more and more products become an exchange of information *about* the user for information *for* the user – e.g., user location for navigation and weather forecast, past purchases for effective recommendations, etc. – product managers and developers must decide what information they will collect, where they will store it, how they will protect it, and whether they will share it with other commercial entities. Thus developers may be seen as mediators whose daily work involves translating cultural values into material products (Ribak, 2019). The challenges to privacy that they identify, and the solutions they design, both mediate and are mediated by the web of norms and practices of which they are a part.

Studies suggest that developers do not prioritize privacy, and are not particularly critical or reflective about the privacy implications of their work (Balebako et al., 2014; Boyd, 2021; Boyd & Shilton, 2021; Hadar et al., 2017; Jørgensen, 2018; Li et al., 2022; Senarath & Arachchilage, 2018). The proposed presentation is designed to contribute to this emergent (rather technical) body of work in two interrelated ways: Conceptually, it draws on digital materialism (Parikka, 2015) to define GitHub[1] (Kitchin, 2017; Prana et al., 2019) as a platform in which culturally diverse open-code developers meet other developers, encounter practical and ethical problems and solutions, learn from one another and collaborate in joint projects. Methodologically, it is attentive to the ways in which developers consider the design of privacy amongst themselves. Following Green and Shilton (2017; Shilton & Green, 2018; Tahaei et al., 2020), it attempts to learn about developers' privacy challenges by unobtrusively studying their discourse; however, by using automated tools, it can offer a broader view of developers' concerns and practices as they are presented to fellow developers.

[1] https://github.com/. GitHub is the repository of more than 200 million open code projects.

Suggested Citation (APA): Eshkol, K.L. and Ribak, R. (2022, November). ***Privacy discourse in GitHub README files***. Paper presented at AoIR 2022: The 23rd Annual Conference of the Association of Internet Researchers. Dublin, Ireland: AoIR. Retrieved from http://spir.aoir.org.

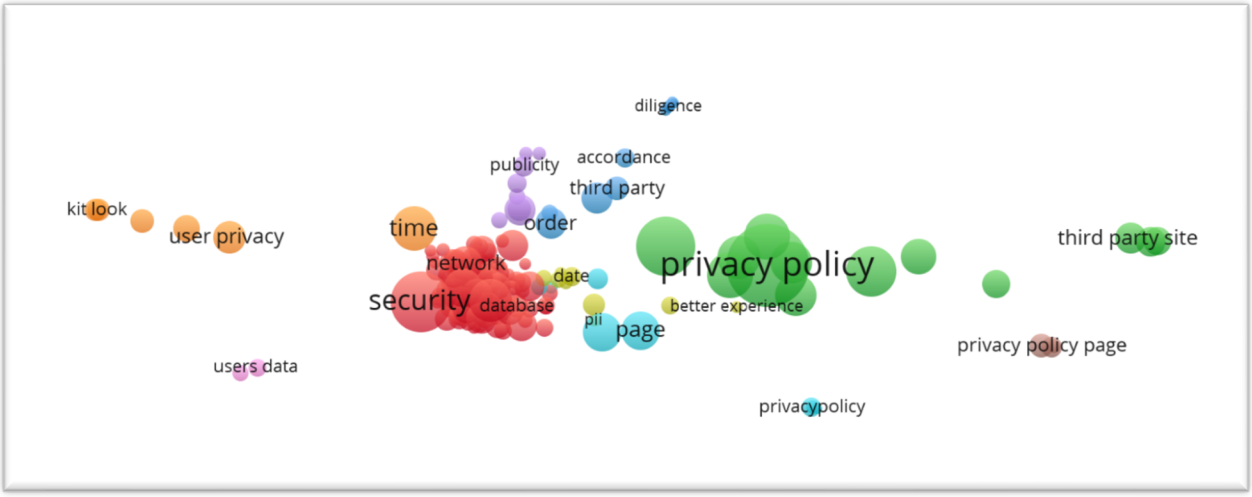
To explore **how developers discuss privacy in GitHub README files,** we adopt a grounded-theory interpretive process that is combined with automated analytic tools based on Natural Language Processing algorithms, designed to make computer languages "understand" human language (Nelson, 2020). We focus on the explanatory files (README) that developers use to account for and essentially promote the software's design and functionality, and where they describe, among other things, how they implement privacy protection.

The first phase of the analysis sought to map developers' privacy discourse. To identify paragraphs that contain information about the developers' privacy practices from within the text, we used GitHub[2] retrieval tools and dedicated software written in Python for information processing.[3] We collected 59,898 README files of software projects on various topics created between 2008-2020 in which the word privacy appears. We saved the relevant paragraph from each file in a separate file used for the automated process. The second phase involved a qualitative discourse analysis of the paragraphs containing the references to privacy. We applied VOSviewer (www.vosviewer.com) for the first phase and Voyant-tools (www.voyant-tools.org) for the second, selected for the level of documentation and functionality they provide.

Preliminary results

The cluster analysis of phase 1 produced nine clusters. The two larger ones represent the main themes in developers' privacy discourse in the README files: One cluster includes 13 words and word pairs (e.g. "external site," "third party"), the most common of which is "privacy policy." The other cluster includes 68 words and word pairs, the most common of which is "security" alongside other, technologically-oriented references (e.g. "algorithm," "anonymity," "communication cloud," "cryptocurrency," "database," "differential privacy," "encryption," "IP address," "machine," "network," "protocol," "transaction").

Figure 1: Clusters of privacy themes in GitHub README files



[2] https://docs.github.com/en/rest/reference/search
[3] https://www.python.org/

To gain insight into developers' privacy approaches (phase 2), we used Voyant-tools to display the key words in their context (KWIC). A qualitative discourse analysis of the paragraphs containing the words found in the two main clusters established that the clusters express two prevalent approaches to maintaining privacy: privacy-by-policy and privacy-by-design (Cavoukian, 2009; Spiekermann and Cranor, 2009), and provided additional insights into developers' privacy discourse.

The initial findings suggest that more than simply two design approaches, the distinction between privacy-by-design and privacy-by-policy discourses is in fact a distinction between discourses that uphold privacy as a value and discourses that regard it as an imposition to comply with. The latter README references consider the policy requirement a legal or corporate obstacle that must be dealt with, and some admit that the application is designed such that displaying the notice and consent button will mislead users to agree to the data collection, thus allowing that app to collect data and use it as it wishes. At the same time, README references that describe privacy-by-design embed solutions in the software code such that it does not allow data misuse, thereby ensuring personal data is used according to the contextual informational norms (Nissenbaum, 2019).

References

Balebako, R., Marsh, A., Lin, J., Hong, J., Cranor, L. F. (2014) The privacy and security behaviors of smartphone app developers. In *Proc. Workshop on Usable Security (USEC'14)*. The Internet Society.

Boyd, K. L. (2021). Datasheets for Datasets help ML Engineers Notice and Understand Ethical Issues in Training Data. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1-27.

Boyd, K. L., & Shilton, K. (2021). Adapting Ethical Sensitivity as a Construct to Study Technology Design Teams. *Proceedings of the ACM on Human-Computer Interaction*, *5*(GROUP), 1-29.

Hadar, I., Hasson, T., Ayalon, O., Toch, E., Birnhack, M., Sherman, S., & Balissa, A. (2018). Privacy by designers: Software developers' privacy mindset. *Empirical Software Engineering*, *23*(1), 259-289.

Jørgensen, R. F. (2018). Framing human rights: exploring storytelling within internet companies. *Information, Communication & Society, 21*(3), 340–355.

Cavoukian, A. (2009). Privacy by design: The 7 foundational principles. *Information and privacy commissioner of Ontario, Canada*, *5*, 12.

Kitchin, R. (2017). Thinking critically about researching algorithms. *Information, Communication & Society* 20.1:14-29.

Li, T., Reiman, K., Agarwal, Y., & Cranor, L. F. (2022). Understanding Challenges for Developers to Create Accurate Privacy Nutrition Labels. CHI '22, New Orleans., LA.

Nelson, L. K. (2020). Computational grounded theory: A methodological framework. *Sociological Methods & Research*, *49*(1), 3-42.

Nissenbaum, H. (2019). Contextual integrity up and down the data food chain. *Theoretical Inquiries in Law*, *20*(1), 221-256.

Parikka, J. (2015). *A geology of media*. U of Minnesota Press.

Prana, G. A. A., Treude, C., Thung, F., Atapattu, T., & Lo, D. (2019). Categorizing the content of GitHub README files. *Empirical Software Engineering*, *24*(3), 1296-1327.

Ribak, R. (2019). Translating privacy: Developer cultures in the global world of practice. *Information, Communication & Society*, 22(6), 838-853.

Senarath, A., & Arachchilage, N. A. G. (2018). Understanding Software Developers' Approach towards Implementing Data Minimization. *USENIX Symposium on Usable Privacy and Security (SOUPS)*, August 12-14, Baltimore MD. *arXiv preprint arXiv:1808.01479*.

Shilton, K. (2013). Values levers: Building ethics into design. *Science, Technology & Human Values*, *38*(3), 374-397.

Shilton, K., & Greene, D. (2019). Linking platforms, practices, and developer ethics: Levers for privacy discourse in mobile application development. *Journal of Business Ethics*, *155*(1), 131-146.

Spiekermann, S., & Cranor, L. F. (2009). Engineering privacy. *IEEE Transactions on software engineering*, *35*(1), 67-82.

Tahaei, M., Vaniea, K., & Saphra, N. (2020, April). Understanding privacy-related questions on Stack Overflow. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-14).