RESEARCH IN PROGRESS

DESIGN OF A DATA ANONYMIZATION TOOL TO ENHANCE SHARING ON AN OPEN DATA PLATFORM

JOHANNES ZEIRINGER,[1] MAXIMILIAN WEBER[2] & STEFAN THALMAN[1]

[1] University of Graz, Graz, Austria.
E-mail: johannes.zeiringer@uni-graz.at, stefan.thalmann@uni-graz.at
[2] Graz University of Technology, Austria.
E-mail: m.weber@student.tugraz.at

Abstract Unintentional disclosure of sensitive data is a critical challenge for many organizations and a serious barrier for open data platforms. Within this research in progress paper, we propose a data anonymization tool to tackle this challenge. The goal of this paper is to elicit design requirements to increase the willingness to share data and collaborate with others on an open data platform. For this purpose, a demonstrator for a data anonymization tool was evaluated within a workshop setting with representatives from companies, science, and public authorities. We found that the willingness to share data can be increased by implementing an anonymization tool and identified further requirements to improve design and to reach the participants' involvement.

Keywords: open data, data privacy, data sharing, knowledge sharing, design science.

University of Maribor Press

DOI https://doi.org/10.18690/um.fov.4.2022.44
ISBN 978-961-286-616-7

1 Introduction and related work

Open data refers to the idea that collected data sets can be viewed, used, or redistributed by, e.g., collaborators on a platform (see: opendatahandbook.org). Looking at, e.g., supply chains (SCs), the volume of collected data increased intensely, by the implementation of advanced digital technologies. This results in SC partners sharing much more data with collaborators (North et al., 2019; Spanaki et al., 2018). Consequently, this increasing exchange of comprehensive data sets leads to problems like privacy issues or even knowledge risks (Spanaki et al., 2018; Zeiringer & Thalmann, 2020). Also, innovation is closely linked to the data exchange between various partners within a SC and also external partners who collaborate in open data platforms and connect public and private stakeholders (Enders et al., 2020; Zeiringer et al., 2022).

The fear to unintentionally disclose critical data and especially critical knowledge is a serious barrier for data, respectively knowledge sharing and for participating in sharing communities (Manhart et al., 2015). Data anonymization is one promising approach to mitigate this fear (Kaiser et al., 2020), and this does not only apply to personal data, but also for product data, process data or machine data, with reference to the previously mentioned SC. Looking at literature, data anonymization is a big field, but when it comes to open data platforms research is scarce and more research needed (Ali-Eldin et al., 2017). The primary goal of this research in progress paper is therefore to identify requirements for a data anonymization tool, to increase the willingness to share data and collaborate with others. For this purpose, the following RQ is addressed:

»What are design requirements for a data anonymization tool to enhance the willingness of data sharing of participants on an open data platform? «

The risk of unintentional knowledge disclosure can lead to knowledge risks. Modern data science methods make it possible to analyze large data sets, and the insights gained in this way could be misused by partners (Zeiringer & Thalmann, 2020). We want to address this problem in the context of open data platforms, which have become a common practice in public sector, but is rather uncommon to firms, especially regarding to the threats of data privacy or strategic reasons (Beno et al., 2017; Enders et al., 2020). To attract entrepreneurs to open data platforms, the

protection of data and the benefits of participation must be ensured. Data anonymization in the context of an open data platform must draw on proven concepts of anonymization in the existing literature, e.g. (Domingo-Ferrer, 2002; Drechsler, 2011; Hundepool, 2012). In the following chapters, the procedure and elaborated design requirements, based on the considerations mentioned above, are presented.

2 Procedure

2.1 Methodology

The problem of unintentional disclosure of sensitive data or resulting knowledge in the context of participating in open data platforms is examined. In a first investigation, the state of research was elaborated within a structured literature review (Zeiringer & Thalmann, 2020). In a further step, an exploratory interview study was conducted, in which various experts were confronted with this problem setting and approaches of dealing, respectively the state of risk management were deduced (Zeiringer & Thalmann, 2021).

Overall we conduct design science research (DSR) (Hevner et al., 2004) as the relevance of our research is directly related to the development of IT artefacts (Peffers et al., 2007). The basic principle in DSR is that research addresses a real world problem by designing an artifact addressing the problem followed by rigorous evaluation showing the impact for practice and theory (Hevner et al., 2004). DSR can be organized according to relevance, design and rigor (Hevner, 2007). The relevance cycle provides requirements from the environment; in our case the idea emerged out of previous workshops and interviews, in which participants have expressed the need for such a tool solution. The rigor cycle (conducted literature review) provided us with the knowledge base to theoretically design such a tool and, last, the design cycle (we present in this paper) aimed at construction and evaluation of the proposed tool (Hevner, 2007).

2.2 Tool design

For elaborating the requirements of a data anonymization tool in an open data environment, a clickable demonstrator was built. Therefore, the *Moqups* platform was used. *Moqups* presents itself as a visual collaboration tool that combines whiteboard, diagram, mockups, and design features in a single, online app. It is web based and easy to use for prototyping (see: https://app.moqups.com). Out of the rigor cycle, multiple types of data anonymization were selected upfront, checked for their feasibility, and implemented in the demonstrator for illustration purposes (see Figure 1, section 2 for selected types).

First, non-perturbative methods were consulted, which replace values of specific description with a less specific description. Two chosen examples are generalization, where individual values of attributes are replaced with a broader category, and suppression, where certain values of the attributes are replaced by, e.g., an asterisk (*) (Hundepool, 2012). Next, perturbative methods were consulted, which distort the data by adding noise, or aggregating values, or generating synthetic data (Hundepool, 2012). First method chosen was additive noise, which replaces the original value with a random added value (Brand, 2002). Second was micro aggregation, which partitions the original dataset into clusters and for each cluster, an aggregation operation is computed and used to replace the original records (Domingo-Ferrer et al., 2002). Lastly, data synthetization refers to data that is artificially created rather than being generated by actual events. Therefore, a model from an original dataset created and by using this model, synthetic data can be generated. This type of data follows the statistical characteristics of the original dataset and does not reveal data points from the original dataset. Synthetic data can either be fully synthetic, which means the entire dataset is replaced, or partially synthetic, which means that only sensitive data is replaced (Drechsler, 2011). The trade-off of privacy and utility is a known impact on the user, when it comes to data sharing, e.g. (Asikis & Pournaras, 2020). Regarding this trade-off it can be said that all methods mentioned above aim to minimize leakage of any kind of sensitive data and try to distort it just enough to keep it useful.

2.3 First demonstrator workshop

The above-mentioned methods for data anonymization aim in two directions: on the one hand, methods are used to anonymize industrial data and, on the other hand, to anonymize personal data. A first demonstration of the tool allowed the participants to get an overview before going into the discussion. Within the workshop, first the overall mission and the goal of the tool was explained to the 15 participants from business, science, and the public sector. All participants had already dealt with data anonymization in advance.

Next the demonstrator was introduced using an exemplary data set containing personal and industrial data. The data to be uploaded is displayed in a preview and the user can already edit the data here (see Figure 1, section 1). By clicking the "Privacy" button, a diagram presents that with absolutely no anonymization, the maximum utility of the data is given. By clicking the "Anonymize" button, a pop-up window appears that illustrates the different types of anonymization and actions (see Figure 1, section 2). The user now can select the approaches to anonymize the data. For each anonymization approach, the user can also change the anonymization attributes, such as the distribution type, the mean and the variance for additive noise, or the number of groups for micro aggregation. Then click "OK" and the data preview shows the first data rows, anonymized as desired.

By clicking the "Privacy" button, the trade-off that takes place between privacy protection and utility of the data (see Figure 1, section 3) is shown. If no more changes are made, the user clicks on continue and can decide in the last step whether the data should now be published or not.

After the demonstration, three rounds, with five people each, were held to discuss the demonstrator, its integration, and further requirements.

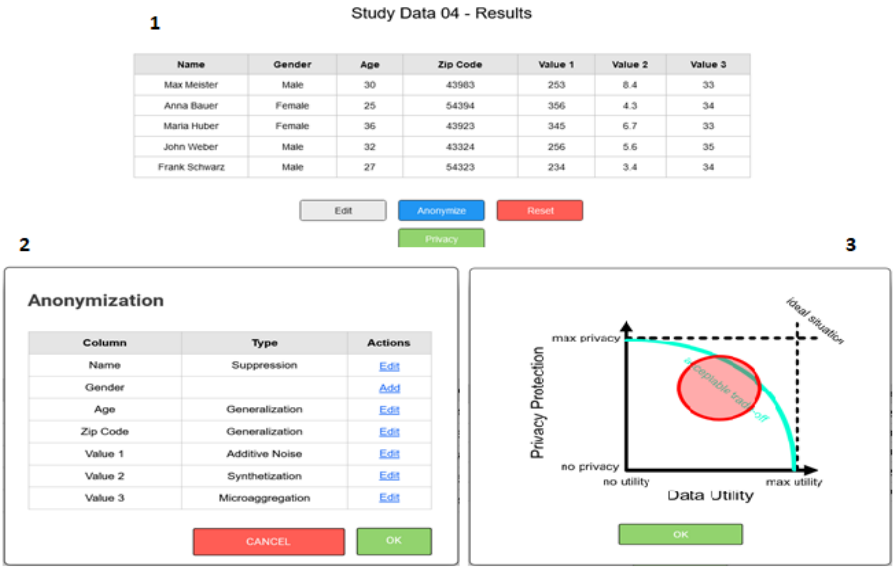

Figure 1: Screenshots of single steps in the demonstrator

2.4 Evaluation and reflections

Four main questions revolved around the integration of a data anonymization tool, for which type of data which anonymization method can be used for, how a tool affects the willingness to share data, and what crucial requirements and ideas for further development came to the participants' attention. The participants highlighted the need for a tool for data anonymization on a potential open data platform and further see such tool as urgent requirement that must be in place before they are willing to share data. Literature recommend certain anonymization techniques, especially for data anonymization prior to sharing: generalization, suppression, permutation, and perturbation (Fung et al., 2010). These techniques were discussed with the participants and considered to be useful and valuable.

Regarding the data types, the need of data anonymization for personal data was often mentioned. Further, the participants also saw uses for industrial data, as well as log data from the internet and mobile networks. These different perspectives were discussed because the workshop participants were from different sectors and dealing with entirely different data. Thus, different types of anonymizations are needed, but the users of the tool must already be trained in advance and know which type of

anonymization fits which type of data. This problem can also be found in the literature and represents a requirement that needs to be addressed in our future research (Hargitai et al., 2018).

Regarding the willingness to share data, certain requirements that clearly need to be in place were emphasized by the participants. First, the benefits must be clear, as it takes time to use. Furthermore, there must be transparency about the processing of data by the tool. Another requirement mentioned was mutuality and reciprocity of data sharing - the willingness to share increases the more people participate. Especially with regard to open data or an open data platform, the person in charge must know exactly which data can be shared without offering conclusions about e.g. internal firm knowledge (Enders et al., 2020). For further development, it was emphasized that use cases for illustration, or tutorials with exciting example data must be made accessible, to make the advantages of the tool, or its application, clear for people. Concluding, the benefit for the user must be clear and comprehensibility about the application possibilities of the different anonymization methods given, it must be free of charge, simple and quick to implement. This is consistent with the criteria listed in acceptance research, such as utility and ease of use (Alexandre et al., 2018).

3 Outlook

This research in progress paper reports on a tool that is designed for data anonymization in open data platforms. The tool aims to reduce or even eliminate data privacy threats and tackle the overall challenge of unintentional disclosure of sensitive data or even resulting knowledge. Thus, users should be able to individually anonymize data before sharing. Within workshops the demonstrator was evaluated, and requirements are elicited.

The availability of a collaborative data platform for open data, which should improve the connectivity of regional partners to international partners is essential. The development and implementation of the demonstrator based on this local open data platform is important for the connection to European initiatives, such as EOSC (see: https://www.eosc.eu/). For future research it is planned to conduct a case study and further build on the demonstrator. The following workshops will address the application of different types of data anonymization and the issue of visualizing data

privacy risks or knowledge risks and make use of user guidance and recommender systems.

References

Alexandre, B., Reynaud, E., Osiurak, F., & Navarro, J. (2018). Acceptance and acceptability criteria: a literature review. Cognition, Technology & Work, 20(2), 165–177. https://doi.org/10.1007/s10111-018-0459-1

Ali-Eldin, A. M. T., Zuiderwijk, A., & Janssen, M. (2017). Opening More Data - A New Privacy Risk Scoring Model for Open Data. In Proceedings of the Seventh International Symposium on Business Modeling and Software Design (pp. 146–154). SCITEPRESS - Science and Technology Publications. https://doi.org/10.5220/0006528301460154

Asikis, T., & Pournaras, E. (2020). Optimization of privacy-utility trade-offs under informational self-determination. Future Generation Computer Systems, 109, 488–499. https://doi.org/10.1016/j.future.2018.07.018

Beno, M., Figl, K., Umbrich, J., & Polleres, A. (2017). Open Data Hopes and Fears: Determining the Barriers of Open Data. In P. Parycek & N. Edelmann (Eds.), Proceedings of the 7th International Conference for E-Democracy and Open Government: Cedem17: Danube University Krems, Krems, Austria : 17-19 May 2017 (pp. 69–81). IEEE. https://doi.org/10.1109/CeDEM.2017.22

Brand, R. (2002). Microdata Protection through Noise Addition. In J. Domingo-Ferrer (Ed.), Lecture Notes in Computer Science: Vol. 2316. Inference control in statistical databases: From theory to practice ; [selected papers (Vol. 2316, pp. 97–116). Springer. https://doi.org/10.1007/3-540-47804-3_8

Domingo-Ferrer, J. (Ed.). (2002). Lecture Notes in Computer Science: Vol. 2316. Inference control in statistical databases: From theory to practice ; [selected papers. Springer. https://doi.org/10.1007/3-540-47804-3

Domingo-Ferrer, J., OGANIAN, A., Torres, A., & MATEO-SANZ, J. M. (2002). ON THE SECURITY OF MICROAGGREGATION WITH INDIVIDUAL RANKING: ANALYTICAL ATTACKS. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05), 477–491. https://doi.org/10.1142/S0218488502001594

Drechsler, J. (2011). Synthetic datasets for statistical disclosure control: Theory and implementation. Lecture notes in statistics: Vol. 201. Springer.

Enders, T., Wolff, C., & Satzger, G. (Eds.) (2020). Knowing What to Share: Selective Revealing in Open Data: European Conference on Information Systems (ECIS), Marrakesch, Marokko, June 15 - 17 2020.

Fung, B. C. M., Wang, K., Chen, R., & Yu, P. S. (2010). Privacy-preserving data publishing. ACM Computing Surveys, 42(4), 1–53. https://doi.org/10.1145/1749603.1749605

Hargitai, V., Shklovski, I., & Wasowski, A. (2018). Going Beyond Obscurity. Proceedings of the ACM on Human-Computer Interaction, 2(CSCW), 1–22. https://doi.org/10.1145/3274335

Hevner, March, Park, & Ram (2004). Design Science in Information Systems Research. MIS Quarterly, 28(1), 75. https://doi.org/10.2307/25148625

Hevner, A. R. (2007). A Three Cycle View of Design Science Research. Scandinavian Journal of Information Systems, 87–92.

Hundepool, A. (2012). Statistical disclosure control. Wiley series in survey methodology. Wiley. https://onlinelibrary.wiley.com/doi/book/10.1002/9781118348239 https://doi.org/10.1002/9781118348239

Kaiser, R., Thalmann, S [Stefan], & Pammer-Schindler, V. (2020). An Investigation of Knowledge Protection Practices in Inter-organisational Collaboration. Protecting Specialised Engineering Knowledge with a Practice Based on Grey-box Modelling. VINE. Advance online publication. https://doi.org/10.1108/VJIKMS-11-2019-0180

Manhart, M., Thalmann, S [Stefan], & Maier, R. (2015). The Ends of Knowledge Sharing in Networks: Using Information Technology to Start Knowledge Protection. In Proceedings of the 23rd European Conference on Information Systems (ECIS), Münster, Germany. https://doi.org/10.18151/7217422

North, K., Carvalho, A. de, Braccini, A., Durst, S., Carvalho, J., Gräslund, K., & Thalmann, S [S.] (2019). Information and knowledge risks in supply chain interactions of SMEs: Proceedings of the 10th International Conference on Practical Knowledge Management, Potsdam, Germany. Lecture notes on Informatics.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. Journal of Management Information Systems, 24(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Spanaki, K., Gürgüç, Z., Adams, R., & Mulligan, C. (2018). Data supply chain (DSC): research synthesis and future directions. International Journal of Production Research, 56(13), 4447–4466. https://doi.org/10.1080/00207543.2017.1399222

Zeiringer, J. P., Durst, S., & Thalmann, S [Stefan] (2022). Show Me What You Do and I Will Tell You Who You Are: A Cluster Typology of Supply Chain Risk Management in SMEs. Journal of Theoretical and Applied Electronic Commerce Research, 17(1), 345–359. https://doi.org/10.3390/jtaer17010018

Zeiringer, J. P., & Thalmann, S [Stefan]. (2020). Knowledge Risks in Digital Supply Chains: A Literature Review. In Proceedings of Wirtschaftsinformatik 2020 (Ed.), WI2020 (pp. 370–385). GITO Verlag.

Zeiringer, J. P., & Thalmann, S [Sefan] (2021). Knowledge sharing and protection in data-centric collaborations: An exploratory study. Knowledge Management Research & Practice, 1–13. https://doi.org/10.1080/14778238.2021.1978886