Framework for a User-friendly Statistical Disclosure Control Tool

Anshika Rawat[1], Mortaza S. Bargh[2], Afshin Amighi[3] and Marijn Janssen[4]
[1,4] Delft University of Technology, Delft, The Netherlands
[2] Research and Documentation Centre, Ministry of Justice and Security, The Hague, The Netherlands
[3] Rotterdam University of Applied Sciences, Rotterdam, The Netherlands
Email: [1] a.rawat-1@student.tudelft.nl [2] m.shoae.bargh@wodc.nl [3] a.amighi@hr.nl [4] m.f.w.h.a.janssen@tudelft.nl

***Abstract*—Public organisations are opening more and more data to increase their transparency. But more often than not, data cannot be opened without modification due to data protection issues. There are Statistical Disclosure Control (SDC) tools that can help to accomplish this. However, these tools are not intuitive for end-users, especially for those who are not deeply familiar with data science. The end-users are mostly public servants who want to open data. Therefore, the main objective of this paper is to develop a framework for a user-friendly tool based on data disclosure methods and usability guidelines for protecting microdata. The developed framework can be used to design and evaluate a tool that can be used by data controllers within the Dutch governmental institutions for anonymising data. This contributes to the sharing of more data with the public or between organisations with a minimised risk of disclosing the identity of the individuals the data is about.**

Keywords–Statistical Disclosure Control; Data Protection.

I. INTRODUCTION

Each year, colossal amounts of data is collected and generated through court proceedings and administrative procedures of the Dutch justice department. This data is generally gathered by many independent organisations that are involved in the Dutch justice system, like the Public Prosecution Service, courts, the Central Fine Collection Agency and the Dutch Police. However, this data is not shared amongst intuitions or made accessible to the public. This is because sharing of such data sets carries an inherent privacy risk. Disclosure of personal data is considered as one of the main threats for data opening. Therefore, only a small proportion of this data is made available to the public to protect the identity of the people the data is about. This impacts the transparency of these organisations because the full potential of these data sets cannot be achieved, as it is paramount to protect the privacy of the respondents of the data because of laws like General Data Protection Regulation (GDPR) [1].

For this reason, the Dutch Ministry of Justice and Security has taken upon itself to enhance its transparency, accountability and efficiency by setting up an open data programme that aims at stimulating the sharing of its data sets with the public or with other organisations [1]. The data sets pertain to information that is gathered by the justice branch of the Dutch government.

It is for this reason that the Research and Documentation Centre (abbreviated as WODC in Dutch) of the Dutch Ministry of Justice and Security is conducting a research on a tool that can be used to anonymize data so that personal data can be protected while maintaining the usefulness of the data. This tool is based on SDC methodologies. SDC refers to methods that try to prevent statistical data from disclosing confidential information about specific respondents, who may be individuals or enterprises [2].

A. Problem Definition

Sizable research has been conducted on appropriate SDC techniques and has been implemented in the form of software tools. Despite the elegance and extensiveness of these solutions, these tools are not widely used by data controllers in public organisations in the Netherlands. The simple reason is that they are complicated, time-consuming to learn and use, and their usage requires a deep understanding of data science. Another reason is that people who implement these techniques, especially statistical agencies, do not widely share, in substantial detail, their knowledge and experience using SDC and about the process of creating safe data with other agencies [3]. This makes it difficult for those organisations who are motivated to implement this solution but are new to the process to get all the relevant information they need to apply these techniques in practice.

B. Research Objective

This project is concerned with the analysis, design and evaluation of a framework for a set of SDC usage guidelines that can be translated into a standardised tool which is user-friendly, time-efficient and intuitive for its intended target user groups. The target user groups of this tool are data controllers within the Dutch public organisations like the Dutch Ministry of Justice and Security.

C. Research Questions

To achieve this, the main objective of this project can be realised by the following five research questions:

1) What is the present state of the SDC tools in use?
2) What are the conceptual SDC methods in practice?
3) How can these conceptual guidelines be transformed into an intuitive tool for the target group?
4) What is the preferred design option, given the needs and skills of the target group and given the organizational setting?
5) How is the preferred design perceived by the target group?

The rest of the paper is structured as follows. In Section 2, we provide a literature overview to introduce key concepts.

 ISBN: 978-1-61208-760-3

Section 3 explains the research methodology to be undertaken. Finally, we conclude the work in Section 4.

II. Literature Overview

In this section, the concept of SDC on data sets relevant for this paper is introduced. In addition to this, the current state of SDC tools is also outlined to show that there is a need for a more user-friendly SDC tool.

A. Conceptual guidelines of SDC

According to the literature, the key guideline of SDC is to protect structured data so that it can be released without giving away the identity of specific individuals or entities [4]. Structured data refers to data that is stored in a structured way, such as a database or spreadsheets. Thus, this technique of anonymising data can be applied to a wide range of data sets, most commonly to microdata and tabular data sets. Microdata refers to data at the individual respondent level. On the other hand, tabular data is aggregate data structured as rows and columns containing information or contributions of a group of respondents. The common output that is offered by national statistical agencies is tabular data [5]. Hutchison & Mitchell distinguish the different data protection methods that can be applied to microdata and tabular data respectively [6]. However, this paper is only concerned with microdata sets and its protections methods. Microdata records include personal data such as direct identifiers which reveal the identity of the person right away. Examples include the name of the person or their address. This is a direct risk to privacy and such identifiers are usually removed or redacted. However, the inherent risk that the data faces is assumed in terms of 'linkage' of sensitive data with identified data [4]. The attributes in the data set that are used in this linkage are termed quasi-identifiers or indirect identifiers. These identifiers do not explicitly reveal the identity of the individual, but are used in combination with other indirect identifiers to re-identify an individual. It is assumed that direct identifiers in the data sets that reveal the identity of the individual such as name, citizenship number, etc, are removed from the data set. However, the literature also indicates that quasi-identifiers cannot simply be removed from the data set. Hundepool provides two reasons for this, first, the data may be required for analysis and, second, the data may already be available to an attacker [4]. Therefore, when designing an SDC tool, the common risks to privacy that can be realised through 'linkage' can be classified into three types [7]. They are:

1) Identity disclosure: It is the foremost risk of re-identification in which the individual can be pinpointed by a specific data entry.
2) Attribute disclosure: This is a primary risk that arises when additional information can be inferred about an individual from the data shared through different data sets or other users.
3) Membership disclosure: A risk in which the attacker is able to determine whether some particular data about an individual may or may not be contained in the data set through data linkage.

Hence, an SDC tool should address the above mentioned risks. On further investigating the literature, it was found that SDC techniques also differ in the way the anonymised data is released. The data release methods can be classified into three types, namely, Public Use Files (PUFs), Scientific Use Files (SUFs) and data made available in a controlled research centre [3]. PUFs are relevant for this topic as this is the data which is made openly accessible to anyone. Because of the public nature of these files, they require protection much larger in the extent to other release types. Therefore, it can be inferred that the design of the tool must have the proper data disclosure capabilities to minimise the risk to user privacy.

However, SDC techniques cannot guarantee the elimination of risk, but they help in reducing the risk to an acceptable level. This is illustrated as a risk-utility trade-off in the SDC process [3]. This trade-off is characterised by the risk of disclosure and the utility of the data for the end-user. The trade-off between the two signifies that, to maximise utility from the data, the risk has to be maximised as well. Figure 1 shows the plot between risk and utility. Zero risk of disclosure is accompanied by the release of no data, whereas data released without disclosure is accompanied by maximum risk. Therefore, SDC techniques have to achieve an optimal point on the risk-utility plot where the maximum utility can be achieved at an acceptable disclosure risk. This is an important concept to remember when comparing and selecting appropriate SDC techniques/methods and choosing the right parameters to mask the data. This understanding should also be reflected in the implementation of the proposed tool. Additionally, the protection methods for microdata should try to reduce the three common risks to privacy by finding the optimal trade-off between risk and quality of data in terms of information loss.

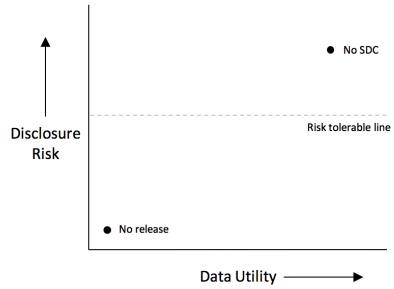

Figure 1. Plot of Risk-Utility Trade-off

The issue of data anonymization is still a complex subject which requires expertise to translate the necessary guidelines for achieving a simplified tool to the point that it can be used by a target group with minimal knowledge of statistics and data science. In the Netherlands, there is no current infrastructure to sanitise confidential data to make it available for external use. Often, data is not released at all and stored away only for authorised personnel to see. This implies that the full potential of this data is never realised, as it cannot be shared for use for conducting research or for other scientific purposes due to GDPR violation. Thus, there is a need to develop a

 ISBN: 978-1-61208-760-3

simple-to-use tool or at least a software environment that can harness some basic data anonymisation technique so that data controllers can adopt some minimal practices of sanitizing data and release them to the public.

This brings us to the question of how can the complexity of SDC techniques be reduced or hidden by an agreeable user-interface? One way to reduce the complexity of the issues is to explore already existing tools that can be simplified so as to present the data controllers with a less than perfect solution, but still something that can tentatively be adopted by them. A simple-to-use interface can guide users through complex data protection methods. Prasser & Kohlmayer suggest that data controllers follow the onion layer principle in which data is protected in a series of multiple layers as there is no single method to protect data sets [7]. These steps including legal agreements, as well as collecting and storing that data, which is absolutely necessary. These measures are already being followed because of regulations like GDPR and by executing a Data Protection Impact Assessment (DPIA).

B. SDC Tools

There are quite a few SDC tools that are available, notably μ-Argus [4], sdcMicro [8] and ARX [9]. Multiple comparative studies have highlighted the capabilities of these tools when compared to each other. The literature shows that tools like μ-Argus implement a broad range of SDC methodologies. On the downside, this tool is a closed-source application which is no longer in active development. SdcMicro is an additional package for existing statistical software. It provides elementary anonymization techniques and has limited support to other specific data transformations. Prasser et al. analyse that the limitation of these tools relate to scalability issues when handling large data sets, incomplete support of privacy criteria and methods of data transformation and, most significantly, these tools require complex configuration by Information Technology (IT) experts [9].

In contrast, ARX is an open-source software which addresses the issues that other tools lack. It provides a broad range of efficient data anonymization methods along with a cross-platform user-interface. This graphical front-end provides multiple perspectives for configuring, exploring and analysing the data. But the main reason ARX stands out is that the framework is designed to prevent tight coupling of subsystems and ensure extensibility. This has helped the developers of ARX provide a stand-alone software library with a public Application Programming Interface (API), which can be used for integrating with other systems.

Even though ARX is the preferred tool of choice, it is not widely used for the simple reason that its many functionalities make this tool quite complex. The extensive features provided by ARX can be quite overwhelming to a user with basic knowledge of this field. The many terminologies used in the graphical front-end of ARX might be unknown to a common user, making this tool not intuitive. ARX also provides a user manual guide and an in-built setup assistant (wizard), but it does not help in reducing its complexity. Gould & Lewis suggest that the different components of a software – operating environment, user platform and reference manuals or materials usually fail to interact cohesively to create a conception that the user eventually deals with, as these components are designed separately [10].

As a result, there is a need to design and implement a tool which would be easier to learn and adopt. The findings suggest that sdcMicro and μ-Argus do not provide the flexibility or the scalability that can be used as a potential development environment for a new user-friendly implementation of an SDC tool. Instead, the availability of the extensive APIs of ARX coupled with its continuous development, testing and documentation due to active development makes this tool ideal for developing a tool that harvests the potential of the ARX tool itself, but in an environment that may be more suited to data controllers and their levels of SDC related knowledge.

C. User Interface Design

Designing a user interface for a software often involves a considerable investment of time and effort which can be reduced by adhering to previously established design guidelines [11]. These guidelines can serve as a starting point for establishing software requirements and development of the framework. Most of these guidelines explore aspects of the user interface design on data entry, data display and Human-Computer Interaction (HCI). However, not all guidelines can be applied to the design and have to be filtered for tailoring the framework design to fit the needs of the target group.

Another technique that can be used for capturing and describing the functional requirements of a software tool is use-case modelling [12]. Use-cases describe all those scenarios in which a user can interact with a system [13]. Writing effective use-cases can help in realising the goals of the different stakeholders. It can also lead to stakeholder driven requirement analysis taking into account the possibility of conflicting requirements [14].

Moreover, Morris & Dillon argue in their paper that developers can gather inputs on user perception of the usefulness or ease of use of the system based on preliminary designs of software tools [15]. The paper suggests that these early formulations of user perception of a system have an influence on whether users will actually use that system [15]. The literature also suggests that capturing predictive measures of user acceptance, even before the user has an opportunity to interact with the software, can lead to correlations between perceived usefulness and eventual user adoption of the software [16].

In conclusion, guidelines and requirement analysis can help in formulating a preliminary design of the framework. This can result in detecting user perceptions early in the design cycle, leading to reduced cost and effective time management.

III. Research Methodology

The nature of this project requires an understanding of SDC methodologies and their architecture so that they can be translated into design guidelines for the proposed framework. Additionally, the ARX tool has to be evaluated for the comprehension of its capabilities and how those capabilities

 ISBN: 978-1-61208-760-3

can be exploited so that the framework can be built upon features that can eventually result in the implementation of a new tool. Design guidelines for a user interface software have to be formalised in conjunction with the goals of the stakeholders to realise the requirements of the framework. Furthermore, use-case modelling and qualitative data analyses can lead to a final design. Lastly, the design has to be evaluated to map it against the goals of all the stakeholders. The five research questions mentioned previously can help in achieving a generalised framework for an SDC tool.

Research questions 1 and 2 can be answered by conducting a literature study that can help in understanding the present state of the current SDC technologies, methods and their required guidelines. By doing this, shortcomings of the current state of the art can be understood and the most relevant ones can be handled in the framework. The literature study can also help in making an informed decision on the models and methods of data anonymization that should be part of the proposed framework. Thus, the first two research questions can help in formulating generalised SDC guidelines. The research also calls upon investigating into HCI, which happens through the user interface design. This can help in further shaping the design guidelines and features of a primary tool and translating it into an intuitive software design. Literature study can further help in answering research question 3. The literature can also reveal insights into user perception and behaviour for the early adoption of the tool. Thus, the first three questions can help in defining a set of guidelines and characteristics of the framework.

Next, to further understand the target group, a qualitative analysis research method approach will be undertaken to understand the requirements of the tool in an organisational setting. This involves conducting semi-structured interviews consisting of open-ended questions with the target group. Such type of interviews can be successful in delving deeper into the content matter which would have otherwise been difficult in structured or focus interviews [17]. The interviews will be conducted with individuals from target organisations within the Dutch public institutions like WODC, Dutch Police, etc.

The guidelines can then be finalised using the literature study and insights gained during the qualitative analysis. Once the guidelines have been agreed upon, requirement engineering can be done via use-case modelling. Next, a framework can be designed for the SDC tool, thereby answering research question 4.

Lastly, the design of the framework has to be validated by the target user-group. A mock-up of the new tool can be made using the framework to see how it is perceived by the target users. The research can explore the design and evaluation of the tool by applying different techniques such as heuristic evaluation, usability testing, use of guidelines and the cognitive walk-through method [18]. Heuristic evaluation and usability testing require the involvement of User Interface (UI) specialists. This is not feasible for this project. Therefore, the use of guidelines and cognitive walk-through techniques can help in finding serious as well as general problems with the usability of the tool without the need for a UI specialist. Thus, the final research question can be answered by conducting a survey and by applying the previously mentioned usability techniques. The result of the survey should be analysed to accommodate the recommendations to make improvements in the framework.

IV. Conclusion

In this position paper, we have identified a problem in practice regarding data disclosure tools and how they are not used by data controllers in public organisations in the Netherlands. We proposed a solution direction to address this problem by designing a framework for a simpler and more user-friendly implementation of an SDC tool.

References

[1] M. S. Bargh, R. Meijer, and M. Vink, "On statistical disclosure control technologies: For enabling personal data protection in open dara settings," Research and Documentation Center WODC, The Hague, The Netherlands, Tech. Rep., Cahier 2018-20.

[2] J. Domingo-Ferrer and V. Torra, "Disclosure risk assessment in statistical data protection," Journal of Computational and Applied Mathematics, vol. 164, 2004, pp. 285–293.

[3] T. Benschop, C. Machingauta, and M. Welch, "Statistical Disclosure Control: A Practice Guide," 2019.

[4] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, R. Lenz, J. Naylor, and P. De Wolf, "Handbook on Statistical Disclosure Control," ESSnet on Statistical Disclosure Control, 2010.

[5] J. Domingo-Feffer, A. Oganian, and V. Torra, "Information-theoretic disclosure risk measures in statistical disclosure control of tabular data," in Proceedings of the International Conference on Scientific and Statistical Database Management (SSDBM), Jan. 2002, pp. 227–231.

[6] D. Hutchison and J. C. Mitchell, "Privacy in Statistical Databases," in Proceedings of the CENEX-SDC Project International Conference, 2006.

[7] F. Prasser and F. Kohlmayer, Putting Statistical Disclosure Control into Practice: The ARX Data Anonymization Tool, 2015, chapter 6, pp. 111–148, in book, Gkoulalas-Divanis, A. and Loukides, G. (Eds.), Medical Data Privacy Handbook, ISBN: 978-3-319-23633-9.

[8] M. Templ, "Statistical disclosure control for microdata using the r-package sdcmicro," Transactions on Data Privacy, vol. 1, no. 2, Aug. 2008, pp. 67–85.

[9] F. Prasser, F. Kohlmayer, R. LautenschlÃger, and K. A. Kuhn, "ARX–A Comprehensive Tool for Anonymizing Biomedical Data," in Proceedings of the Annual Symposium / AMIA Symposium, 2014, pp. 984–993.

[10] J. D. Gould and C. Lewis, "Designing for usability-key principles and what designers think," in Proceedings of the Conference on Human Factors in Computing Systems, vol. 28, 1983, pp. 50–53.

[11] S. L. Smith and J. N. Mosier, "Guidelines for designing user interface software," 1986.

[12] B. Anda, H. Dreiem, D. I. K. Sjøberg, and M. Jørgensen, "Estimating Software Development Effort Based on Use Cases — Experiences from Industry," in The Unified Modeling Language. Modeling Languages, Concepts, and Tools, M. Gogolla and C. Kobryn, Eds. Springer Berlin Heidelberg, 2001, pp. 487–502.

[13] D. Rosenberg and K. Scott, Use case driven object modeling with UML. Springer, 1999.

[14] B. G. Cameron, E. F. Crawley, G. Loureiro, and E. S. Rebentisch, "Value flow mapping: Using networks to inform stakeholder analysis," Acta Astronautica, vol. 62, 2008, pp. 324–333.

[15] M. G. Morris and A. Dillon, "How user perceptions influence software use," IEEE Software, vol. 14, no. 4, July 1997, pp. 58–65.

 ISBN: 978-1-61208-760-3

[16] F. D. Davis and V. Venkatesh, “Toward preprototype user acceptance testing of new information systems: implications for software project management,” IEEE Transactions on Engineering Management, vol. 51, no. 1, Feb 2004, pp. 31–46.

[17] B. DiCicco-Bloom and B. F. Crabtree, “The qualitative research interview,” Medical Education, vol. 40, no. 4, 2006, pp. 314–321.

[18] R. Jeffries, J. R. Miller, C. Wharton, and K. Uyeda, “User interface evaluation in the real world: a comparison of four techniques,” in Proceedings of the SIGCHI conference on Human factors in computing systems, 1991, pp. 119–124.

 ISBN: 978-1-61208-760-3