Challenging the robustness of OGD de-identification rules through a hackathon

Auriane Marmier[1] and Tobias Mettler[1]

[1]*Swiss Graduate School of Public Administration, University of Lausanne, Lausanne, Switzerland, {auriane.marmier,tobias.mettler}@unil.ch*

Abstract: With the emergence of the notion of "open innovation", public organisations are currently undergoing a transformation process. Particularly driven by the idea of open government, the release of data that has been produced and financed by public funds has increased and with it, the risk associated to the publication of sensitive or personal information about citizens. Although the diffusion of open government data (OGD) might be beneficial for the private sector, the disclosure of such data might engender several risks, which could affect an individual's privacy. In order to avoid this issue, governments worldwide have started to protect the privacy of individuals by applying de-identification rules. However, de-identification is not risk-free. If the de-identified data does not provide sufficient robustness, re-identification (or re-construction) of personal information is possible. In this paper, we describe a practical approach to examine OGD de-identification rules.

Keywords: De-identification, Open Government Data, Action Design Research, Hackathon

1. Introduction

Public organisations have released a large amount of open government data (OGD) in the past years. With the emergence of open innovation, the interest for such data have grown steadily and require public organisations to rethink their mode of governance. With the progressing digitalization of the public administration, it has become easier to exploit ever-larger amounts of public data that is published on open government platforms or government websites. This data may reveal important information about diseases, financial situation, or consumption habits of an individual. The disclosure of such data may engender several risks, which may lead to extremely negative consequences (e.g. identity theft, fraud, job or reputation loss of an individual) (Erdem and Prada 2011) (Benitez and Malin 2010) and consequently may affect individuals' privacy (Paspatis, Tsohou et al. 2017). In order to avoid this issue, governments worldwide have started to care more about de-identification by implementing specific techniques to ensure that sensitive information is not disclosed to third parties. However, de-identified data has a significant risk of re-identification if these techniques are not sufficiently robust.

In this paper, we therefore describe an approach related to the following research question: How to evaluate the robustness of OGD de-identification rules? For that matter, we propose to follow

Action Design Research (ADR) (Sein, Henfridsson et al. 2011), a pragmatic research method that links practitioners and researchers to build a concrete solution (i.e. in our case de-identification rules for OGD) and evaluate the robustness of these rules during a hackathon. By applying ADR to build and evaluate our artefact, we consider the specific needs of public organisations, create a practical solution as well as simultaneously test it in a safe place, under real conditions (i.e. a hackathon). In view of challenges that appear when we contrast data privacy theories to a real-world problem (i.e. re-identification of personal information) (Sein, Henfridsson et al. 2011) hackathons provide to public organisations a sort of think tank to observe and evaluate the suitability of developed techniques.

This paper is organised as follows. First, we explore the current state of discussion concerning de-identify technics and re-identify risks in the literature. Then, we explain the process of the ADR method, followed by a complete description of the design proposed to public organisations to build and test de-identification rules (see Fig. 1). We finally discuss the reasons that a hackathon seems to be an appropriate solution for assessing the robustness of de-identification.

2. Background

2.1. Data protection and de-identification

In the perspective of ensuring private life, limiting risks but still providing analytical utility for researchers, Erdem and Prada (2011) note that authorities must be responsible to assure data confidentiality. According to El Emam (2016), open data initiatives should not release personal information. Joo, Yoon et al. (2018) argue that for this reason, OGD programs need robust strategies to manage the publication of personal and sensitive information. Many public organisations have adopted diverse technics to mitigate risks of data privacy. Although in the literature the technics used for de-identified varies (Benitez and Malin 2010, El Emam, Arbuckle et al. 2012, Scaiano, Middleton et al. 2016) researchers agreed that de-identification tends to reduce the risk associated to the publication of sensitive and personal information (Benitez and Malin 2010, U.S. Department of Health & Human Service 2010, Scaiano, Middleton et al. 2016). These are appropriate technics to secure citizen data (Iverson and Davis 2007). Joo, Yoon et al. (2018) note that after being de-identified, such type of data can be freely used as any other OGD.

Among the various de-identification techniques (Finkle , El Emam 2016, Office for Government Policy Coordination 2016, Simson G 2016), the utilisation of de-identification rules on public data sets is one of the most used alternatives (Joo, Yoon et al. 2018). That technique consists of transforming sensitive data elements by applying dedicated rules (El Emam and Arbuckle 2013). Experts recognize two types of sensitive data elements: direct identifiers and quasi-identifiers (or an indirect identifier) (Willenborg and De Waal 1996, Duncan, Elliot et al. 2011, El Emam and Arbuckle 2013, Scaiano, Middleton et al. 2016). In the literature, variables that can be used alone to uniquely identify individuals are defined as a direct identifier of a data set (e.g. social security number, telephone number, voter identification number, etc.) (Benitez and Malin 2010, El Emam and Arbuckle 2013, Czajka, Schneider et al. 2014). Guidelines also treat name or email as direct identifiers

because on a given data set there is generally only one individual that has this name or email address (El Emam and Arbuckle 2013). On the other hand, quasi-identifiers are variables that represent the contextual information about individuals that can be used to indirectly identify them (e.g. gender, date of birth, age, geographical information, zip codes, spoken language, ethnic origin etc.) (Benitez and Malin 2010, El Emam, Arbuckle et al. 2012, El Emam and Arbuckle 2013, Scaiano, Middleton et al. 2016). They are often used in combination (e.g. ethnicity, birthdate, and geographical location) (El Emam and Arbuckle 2013).

2.2. Re-identification risks

While there are many ways to de-identify data sets, various scholars have demonstrated that transformed data can be easily re-identified (Porter and Tech. 2008, De Montjoye, Radaelli et al. 2015, Financial Times 2016, Paspatis, Tsohou et al. 2017). The combination of a voter list (providing information such as birthdate, gender and residential ZIP code) with hospital discharge records have been led to the re-identification of the medical record of the Massachusetts governor (Sweeney 2000). Culnane, Rubinstein et al. (2017) demonstrate that limited information is sufficient in the process of an individual's re-identification. Similarly, Paspatis, Tsohou et al. (2017) show that sensitive and personal information is present everywhere on the web and the sharing of only three of them is sufficient to re-identify an individual.

We note that de-identification guidelines only recommend risk assessments of re-identification but not to test in real condition if these de-identification rules can be breached having a certain background, situational information. In the context of de-identification, researchers note that often risk evaluation approaches are limited (Meystre, Friedlin et al. 2010). Benitez and Malin (2010) developed two risk estimation metrics that allow them to evaluate the probability of re-identification for each record in a published data set. The study conducted by Scaiano, Middleton et al. (2016) shows that the metrics used for the evaluation do not consider the risks associated with the release of data (i.e. they only include data that needs to be treated, without taking into account its environment). In other words, practitioners traditionally de-identify data as if they were independent, excluding the impact of the real world. In general, governmental guidelines (Finkle , Canadian Institute for Health Information 2010, National Institute of Standards and Technology 2012, Office for Government Policy Coordination 2016) recommend conducting risk assessment procedures before the publication of data. They usually propose qualitative or quantitative approaches. Qualitative approaches base the risk assessment from a set of questions such as the number of data, the level of security control, the type of data etc. while quantitative approaches measure the risk of re-ID as the probability that someone will find the correct identity of a single individual. Simson G (2016) recommends evaluating rules through a software and Finkle () advices the utilisation of a motivated intruder test (i.e. data experts try to re-identify data sets). We note that the rules are rarely subject to the conditions they will meet once published (e.g. hacker, media, researchers etc.). Through this study, we would like to see if by placing the de-identified test data in a real and controlled environment (hackathon) would allow us to better understand the risks associated with re-identification.

3. Methodology

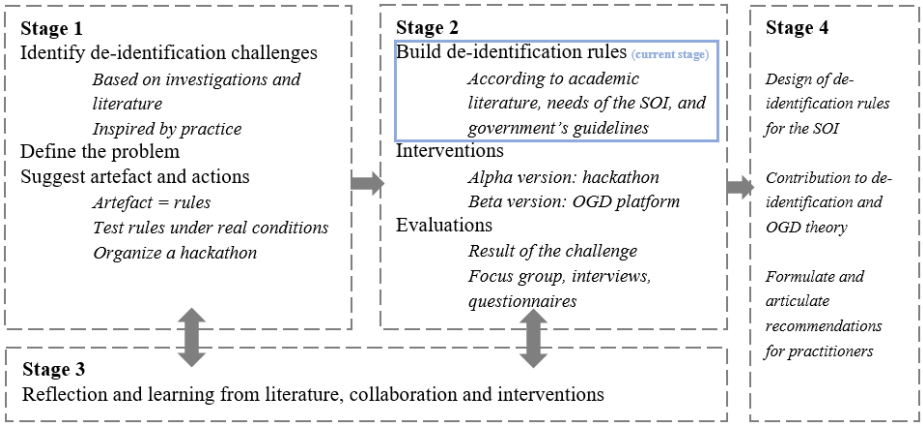
According to Petersson and Lundberg (2016), the Action Design Research (ADR), introduced by Sein, Henfridsson et al. (2011), can be used to foster new ideas for the creation of new technical solutions as well as organizational learning. ADR is composed of four stages: 1-Problem formulation (i.e. specify a concrete problem), 2-Building, Intervention and Evaluation (i.e. finding and evaluating solutions in an iterative process between practitioners and researchers), 3-Reflecting and Learning (i.e. after each iteration add recommendations for an artefact) and 4-Formalisation of Learning (i.e. give final recommendation) (Sein, Henfridsson et al. 2011), ADR provides researchers and practitioners with a systematic approach for planning a participative intervention, as we plan to do with our hackathon for testing de-identification rules (Sein, Henfridsson et al. 2011). The use of ADR allows researchers to learn from the intervention, optimise the artefact, find applicable recommendations for organisations or resolve problematic situations (Baskerville and Wood-Harper 1998, Von Alan, March et al. 2004, Sein, Henfridsson et al. 2011, Crow and Shangraw 2016). An artefact often emerges from the intervention of researchers to solve organisational challenges (Sein, Henfridsson et al. 2011). This artefact typically represents material and organisational features emerging from design, use, and ongoing refinement in a project (Sein, Henfridsson et al. 2011). In our case, the artefact is the de-identification rule set which we develop in collaboration with academic researchers and the Organisational and Informatics Service (SOI) of a major city in Switzerland.

4. Application of ADR

In the problem formulation stage, we examined the challenges posed by the de-identification process. We have discovered that several actors appear in that process: the researchers, the SOI, (i.e. IT department in municipal level), the housing department (i.e. the owners of OGD), citizens (i.e. they are directly affected by data disclosure risks) and OGD end-users. One of the goals of the SOI is to « facilitate » the publication of the city-owned data. It is its responsibility to ensure that the legal base about anonymization is respected when data is being published in open access. We apply the de-identification process on the housing department data mostly because the two services (i.e. SOI and housing department) maintain a good working relationship but also because the city for this study is actually engaged in sustainable development policy. Data such as energy consumption of an urban building or water consumption of citizen appears to be the most attractive option in terms of data valorisation and therefore potentially the most used for re-identification. Together with the SOI and the housing department, we form the ADR team. Based on the literature, practice guidelines, as well as personal meetings and participant observation of the SOI and the housing department, the ADR team has shaped the concrete problem: the individual disclosure risk assessments is weak and misunderstood. We concluded from our examination of the literature that there is a lack of practical approaches for re-identification risk assessment. Consequently, we formulated our research question: How to evaluate the robustness of OGD de-identification rules? In order to answer this question, the ADR team recommended organising a hackathon, to test the rules under a real and controlled situation.

We are (i.e. the researchers) currently in the initial phase of stage two (see Fig. 1). Based on the needs of the SOI and the housing department, the gaps described in the academic literature as well as other government recommendations we built the de-identification rules (Finkle). As recommended in existing guidelines (U.S. Department of Health & Human Service 2010, Office for Government Policy Coordination 2016, Simson G 2016), we first sort out and selected the most relevant data to be used in such a process (e.g. application to control its energy consumption or to reduce water use). In accordance with the SOI, we identified the data used for potential individual disclosure, sensitive data and data considered to be risk-free. The latter was excluded from any de-identification. We also decided to exclude sensitive data (define by the European General Data Protection Regulation (GDPR) list), in order to limit ethical risks (Regulation 2016). As the characteristic of data defines the type of de-identification, we then together determined among the data remaining direct identifiers and quasi-identifiers (El Emam 2016). Then, for each group, we determined the characteristics of the data (e.g. numerical or geographical data, date and times, unstructured text etc.) and finally chose the most suitable rule.

Figure 1: Action Design Research applied to de-identification rules process



The hackathon, the intervention that will serve to test the alpha version of our artefact, has not yet taken place. We will challenge the de-identified data during the hackathon as well as continue our observations, interviews and survey in order to still reflect and learn about de-identification challenges and artefacts. End users of OGD will be invited to explore the de-identified data and imagine re-identification attacks. Insights from this step will help us to develop a robust rule set for a specific domain that will then be published on an OGD platform.

Finally, the formalisation of learning stage will take the form of final de-identification rules and recommendations addressed to the SOI about requirements and needs in publishing OGD safely.

5. Discussion

Heeney, Hawkins et al. (2011) argue that privacy risk assessments need to consider the whole data environment and not only a single data set. According to El Emam and Arbuckle (2013), the techniques used to achieve de-identification should not be separated from their context and should be evaluated on a larger scale, including the environment of the data as well as be implemented

across varied sources (Meystre, Friedlin et al. 2010). In the context of de-identification, we note that privacy risk assessments often ignore these recommendations. Public organisations rarely test de-identification rules in a final data environment (i.e. OGD platforms) and consequently, few are confronted with real-world challenges. In the literature, scholars have started to explore innovative contests as appropriate for proposing practical solutions in de-identification challenges (Möslein and Bansemir 2011, Hjalmarsson, Johannesson et al. 2014, Juell-Skielse, Hjalmarsson et al. 2014, Juell-Skielse, Hjalmarsson et al. 2014), but there are a limited number of solutions proposed to integrate environment impact on de-identification rules (Dinter and Kollwitz 2016). By linking the actors in a situational environment with the de-identification techniques and de-identified data sets, we believe to perform a participative intervention such as those presented by Hjalmarsson, Johannesson et al. (2014) (e.g. collaboration, innovative workshops (Möslein and Bansemir 2011), online setting (Möslein and Bansemir 2011) or hackathons), in order to challenge the robustness of de-identification, which to our view is one of the most important steps in the OGD publication process. For Dinter and Kollwitz (2016), participative innovations constitute a suitable method for innovating with OGD. Owing to the emergence of open innovation, we are convince that the hackathon is suitable to meet de-identification challenges. Often rewarded with a price, hackathons usually regroup participants as teams, over a short period (e.g. 24 to 48 hours) with the aim to resolve dedicated issues. Teams may be composed of highly qualified individuals as well as people curious and interested in innovations. In general, teams work on diverse challenges such as data processing or any other computing problem (Briscoe and Mulligan 2014). Such events make it possible to estimate locally, from a different point of view, the risks of re-identification prior to sharing data on open platforms (Benitez and Malin 2010). We see the hackathon as a sort of laboratory to test the robustness of the rules. It represents a safe place for authorities because they released data in a restricted area to controlled groups of people, surrounded by experts of the domain and citizen representatives. It allows the ADR team to gather activities such as operate in the organisation, improve de-identified rules and evaluate them concurrently. By putting our de-identification rules in real-life conditions while remaining in a protected environment, we are then able to observe them and understand their weakness. Furthermore, the hackathons give the opportunities to the SOI to have an overview of the housing department data that interests the most participants, better targeting their needs and therefore adapting to the de-identification levels required. It is not only a question of testing an artefact but also for understanding, the issues related to the data that will soon be published.

References

Baskerville, R., & Wood-Harper, A. (1998). *Diversity in information systems action research methods*. European Journal of information systems 7(2): 90-107.

Benitez, K., & Malin, B. (2010). *Evaluating re-identification risks with respect to the HIPAA privacy rule*. Journal of the American Medical Informatics Association 17(2): 169-177.

Briscoe, G., & Mulligan, C. (2014). *Digital innovation: The hackathon phenomenon*. Arts and Humanities Research Council, the European Regional Development Fund.

Canadian Institute for Health Information (2010). "*Best Practice*" *Guidelines for managing the disclosure of De-Identified Health Information*. Canadian Institute for Health Information, Health System Use Technical Advisory Committee Data De-Identification Working Group.

Crow, M., & Shangraw, J. (2016). *Revisiting "Public Administration as a Design Science" for the Twenty-First Century Public University*. Public Administration Review 76(5): 762-763.

Culnane, C., et al. (2017). *Health Data in an Open World*. Retrieved 2019/02/16, from https://arxiv.org/abs/1712.05627.

Czajka, J., et al. (2014). *Minimizing disclosure risk in HHS open data initiatives*. Mathematica Policy Research, Washington DC.

De Montjoye, Y., et al. (2015). *Unique in the shopping mall: On the reidentifiability of credit card metadata*. Science 347(6221): 536-539.

Dinter, B., & Kollwitz, C. (2016). *Towards a framework for open data related innovation contests*. Pre-ICIS SIGDSA/IFIP WG8. 3 Symposium. Dublin, Ireland.

Duncan, G., et al. (2011). *Statistical confidentiality: principles and practice*. New York, Springer-Verlag

El Emam, K. (2016). *A de-identification protocol for open data*. Retrieved 2019/02/14, from https://iapp.org/news/a/a-de-identification-protocol-for-open-data/.

El Emam, K., & Arbuckle, L. (2013). *Anonymizing health data: case studies and methods to get you started*. Sebastopol, O'Reilly Media, Inc.

El Emam, K., et al. (2012). *De-identification methods for open health data: the case of the Heritage Health Prize claims dataset*. Journal of medical Internet research 14(1): e33(1).

Erdem, E., & Prada, S. (2011). *Creation of public use files: lessons learned from the comparative effectiveness research public use files data pilot project*. JSM Proceeding (2011). Miami Beach, Florida.

European Parliament and Council of The European Union (2016). *Regulation (EU) 2016/679 of the European Parliament and of The Council. General Data Protection Regulation*. Retrieved 2019/02/03, from https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1558176381563&uri=CELEX:32016R0679.

Vasagar, J (2016). *Kreditech: A credit check by social media*. Retrieved 2019/02/09, from http://www.ft.com/cms/s/0/12dc4cda-ae59-11e5-b955-1a1d298b6250.html?siteedition=intl#axzz4Ij8hHIA4.

Finkle, E (2016). *Open Data Release Toolkit*. Privacy Edition. Retrieved 2019/02/15, from https://datasf.org/resources/open-data-release-toolkit/.

Heeney, C., et al. (2011). *Assessing the privacy risks of data sharing in genomics*. Public Health Genomics, 14(1): 17-25.

Hevner, A. R., et al. (2004). *Design science in information systems research*. MIS quarterly 28(1): 75-105.

Hjalmarsson, A., et al. (2014). *Beyond innovation contests: A framework of barriers to open innovation of digital services*. ECIS Proceedings. Tel Aviv, Israel.

Iverson, D., & Davis, K. (2007). *System and method of de-identifying data*. Retrieved 2019/03/08, from https://patents.google.com/patent/US20030220927A1/en.

Joo, M., et al. (2018). *De-identification policy and risk distribution framework for securing personal information*. Information Polity 23(1): 1-25.

Juell-Skielse, G., et al. (2014). *Is the public motivated to engage in open data innovation?* In: Janssen M., Scholl H.J., Wimmer M.A., Bannister F. (eds) Electronic Government. EGOV 2014. Lecture Notes in Computer Science, vol 8653. Springer, Berlin Heidelberg.

Juell-Skielse, G., et al. (2014). *Contests as innovation intermediaries in open data markets*. Information Polity 19(3, 4): 247-262.

Meystre, S., et al. (2010). *Automatic de-identification of textual documents in the electronic health record: a review of recent research*. BMC Medical Research Methodology 10(1): 70.

Möslein, K. M., & Bansemir, B. (2011). *Strategic open innovation: basics, actors, tools and tensions*. In: Hülsmann M., Pfeffermann N. (eds) Strategies and Communications for Innovations. Springer, Berlin, Heidelberg.

Office for Government Policy Coordination (2016). *Guidelines for De-identification of Personal Data*. F. S. C. Korea Communications Commission, Ministry of Science, ICT and Future Planning, Ministry of Health and Welfare, Office for Government Policy Coordination, Ministry of Interior.

Paspatis, I., et al. (2017). *Mobile Application Privacy Risks: Viber Users' De-Anonymization Using Public Data*. MCIS Proceeding, Genoa, Italy.

Petersson, A., & Lundberg, J. (2016). *Applying action design research (ADR) to develop concept generation and selection methods*. Procedia CIRP 50: 222-227.

Porter, C. C. (2008). *De-identified data and third party data mining: the risk of re-identification of personal information*. Shidler JL Com. & Tech. 5: 1.

Ronald, S. R. (2012). *Guide for Conducting Risk Assessments*. US Department of Commerce, National Insitut of Standards and Technology. Special Publication (NIST SP) - 800-30.

Scaiano, M., et al. (2016). *A unified framework for evaluating the risk of re-identification of text de-identification tools*. Journal of biomedical informatics 63: 174-183.

Sein, M. K., et al. (2011). *Action Design Research*. MIS quarterly 35 (1) 37-56.

Simson, G. (2016). *De-Identifying Government Datasets*. US Department of Commerce, National Institute of Standards and Technology. Special Publication (NIST SP) - 800-188.

Sweeney, L. (2000). *Simple demographics often identify people uniquely*. Retrieved 2019/02/22, from https://dataprivacylab.org/projects/identifiability/paper1.pdf.

U.S. Department of Health & Human Service (2010). *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule*. Retrieved 2019/02/14, from https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html.

Willenborg, L., & De Waal, T. (1996). *Statistical disclosure control in practice*. Springer-Verlag, New York.

About the Author

Auriane Marmier

Auriane Marmier is a doctoral candidate at the Swiss Graduate School of Public Administration, University of Lausanne, where she writes her dissertation on open government data.

Tobias Mettler

Tobias Mettler is Associate Professor at the Swiss Graduate School of Public Administration, University of Lausanne. His research interests are in the area of design science research, technology adoption, applications of data science, and business models, with a particular focus on public sector innovations.