An interaction model for de-identification of human data held by external custodians

Andrew J. Simmons
Deakin University
Geelong, Australia
a.simmons@deakin.edu.au

Maheswaree Kissoon Curumsing
Deakin University
Geelong, Australia
m.curumsing@deakin.edu.au

Rajesh Vasa
Deakin University
Geelong, Australia
rajesh.vasa@deakin.edu.au

ABSTRACT

Reuse of pre-existing industry datasets for research purposes requires a multi-stakeholder solution that balances the researcher's analysis objectives with the need to engage the industry data custodian, whilst respecting the privacy rights of human data subjects. Current methods place the burden on the data custodian, whom may not be sufficiently trained to fully appreciate the nuances of data de-identification. Through modelling of functional, quality, and emotional goals, we propose a de-identification in the cloud approach whereby the researcher proposes analyses along with the extraction and de-identification operations, while engaging the industry data custodian with secure control over authorising the proposed analyses. We demonstrate our approach through implementation of a de-identification portal for sports club data.

CCS CONCEPTS

• Security and privacy → Usability in security and privacy • Human-centered computing → User models; Heuristic evaluations; Interface design prototyping

KEYWORDS

Anonymization, De-identification, Emotions, Privacy

ACM Reference format:

Andrew J. Simmons, Maheswaree Kissoon Curumsing and Rajesh Vasa. 2018. An interaction model for de-identification of human data held by external custodians. In *Proceedings of the 30th Australian Computer-Human Interaction Conference (OzCHI '18), December 4–7, 2018, Melbourne, VIC, Australia*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3292147.3292207

OzCHI'18, December 4–7, 2018, Melbourne, VIC, Australia

ACM 978-1-4503-6188-0/18/12…$15.00
https://doi.org/10.1145/3292147.3292207

1 INTRODUCTION

Researchers often wish to reuse a pre-existing dataset in a new unforeseen way to investigate a question. As a running example, in this paper we will consider a sports researcher requesting access to a dataset held by a sport club. As obtaining consent of every individual in the dataset to reuse their data is often infeasible, human ethics guidelines permit exemptions if the data custodian de-identifies their dataset and provides it to the researcher in non-identifiable form, i.e. such that no individual can be re-identified [1].

As the data custodian may not be an expert in de-identification techniques, it is important that the de-identification system be learnable, and that it minimize the risk of user errors by the data custodian that could undermine the privacy of participants. Typically, sports club staff would be familiar with business spreadsheet software, such as Microsoft Excel, and remove or substitute identifiable columns such as player names.

However, de-identifying data is a non-trivial operation, as even after obvious identifiers are removed, "quasi-identifiers" [2], such as times, dates, or locations, may still allow re-identifying individuals in the dataset by linking sensitive data to public datasets. Privacy researchers have proposed software tools that automatically distort or generalize quasi-identifiers [3], however use of these tools requires a level of expertise from the user to select an appropriate privacy threshold, ensure that algorithm assumptions are met, and to minimize the destruction of data utility [4]. As sport club staff are under constant time pressure, it is unlikely that they would have time to develop the necessary expertise to apply these tools reliably, and the additional work and uncertainty may cause frustration that undermines the research partnership.

While existing tools for de-identification focus on quality goals or functional requirements, we argue for a solution that also meets the stakeholders' emotion-oriented requirements to ensure the research-industry engagement is successful. Our focus is on the human–computer interactions involved in the de-identification workflow; we abstract over the specific choice of de-identification operation as this is best left to the researcher to decide given the type of dataset and privacy level requirements.

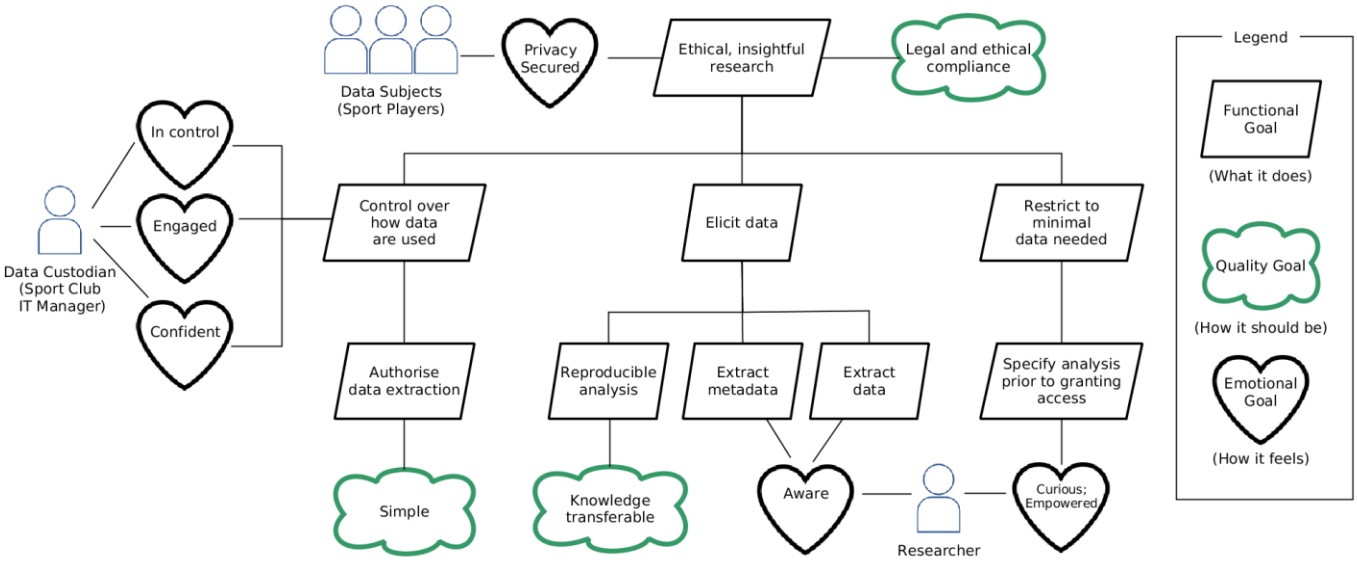

Figure 1: De-identification portal Emotional Goal Model (diagram should be read top to bottom)

2 EMOTIONAL GOAL FRAMEWORK

While functional and quality goals are well-established as part of the software design process, all too often software designers overlook emotional needs of users [5, 6], resulting in an unfulfilling product which fails to gain appropriation by users as part of their workflow [7, 8].

The importance of users' emotional expectations during software design cannot be undermined. In our work, we look at the techniques proposed by [9, 10] to introduce the concept of emotional goals within our software design process.

Specifically, in this paper, we utilize emotional goal modelling to consider the needs of each stakeholder in the design of the de-identification platform. User emotional acceptance of the system is critical to improving data sharing practices, else stakeholders are likely to revert to flawed but culturally engrained [11] data sharing practices, such as substituting names with a randomized code while doing little to prevent the re-identification of individuals via data linkage using quasi-identifiers.

3 MODELLING

3.1 Emotional Goal Model

We break down the functional goals of the system and consider how these impact on the quality and emotional goals of users within Fig. 1.

The overall goal of the de-identification portal is to provide a platform for ethical, insightful research that facilitates reuse of data without compromising the privacy of the data subjects (i.e. the sport players). To achieve this goal, the control over how the data are used must lie with the data custodian (i.e. the sport club) rather than the researcher, as providing the researcher with unrestricted access to the system would be equivalent to transfer of identifiable data without participant consent. On the other hand, to encourage insightful research, the system should promote a mindset of intellectual *curiosity* whereby the researcher feels *empowered* to request (but not necessarily be granted access to) data and propose analyses that fully utilize the detail available in the dataset to gain an *awareness* that is not limited to traditional predefined summary statistics. To meet the goals of all stakeholders, we propose that the researcher should precisely specify the data they need for an analysis by writing a script to perform the extraction. In cases where an analysis requires access to sensitive data, the extraction script should perform the analysis on the sensitive data then de-identify the output to ensure that it is non-identifiable. The data custodian (i.e. the sport club IT manger) should feel *engaged* in the research process and *in control* over authorising the execution of a script so that they can be *confident* the research is protecting the data subjects' (i.e. the players in their club) right to feel that their *privacy is secured*.

3.2 Interaction Model

In Fig. 2 we provide an interaction model for the proposed system that translates the goals into role interactions. At this level we introduce the role of a cloud data portal, an autonomous agent that will mediate the interactions between the data custodian and researcher in a secure manner. To ensure the data custodian remains in control of data access, in our solution they encrypt the data prior to uploading the data to the cloud. The researcher proposes an analysis by uploading a script to the cloud portal and providing a human readable summary for the data custodian. If the data custodian is satisfied that the proposed analysis is respectful of the privacy and rights of the participants, they authorise the cloud portal to perform the analysis proposed by the researcher by providing it with the decryption key. The cloud

portal uses the key to temporarily decrypt the data, runs the analysis script against the raw data, and finally destroys the key after script execution is complete. Upon completion the researcher is notified so that they can perform post-analysis on the results of the extraction script and communicate the results back to the data custodian. This is an iterative process; the first iteration is usually to extract metadata and validate the researcher's assumptions about the dataset. The following iterations deal with extracting data to answer a specific research question, which may prompt subsequent questions.

As the data custodian is *in control* of the authorisation of each phase, this has the additional benefit of keeping them emotionally *engaged* in the research process. As the analysis is run in the cloud, the researcher never sees the raw data nor the decryption key, and thus never has access to re-identifiable data; the researcher is *aware* only of the final output of their analysis, thus *empowering* the researcher to satisfy their feeling of *curiosity* about well-formed questions without revealing details that would compromise the data subjects' privacy.

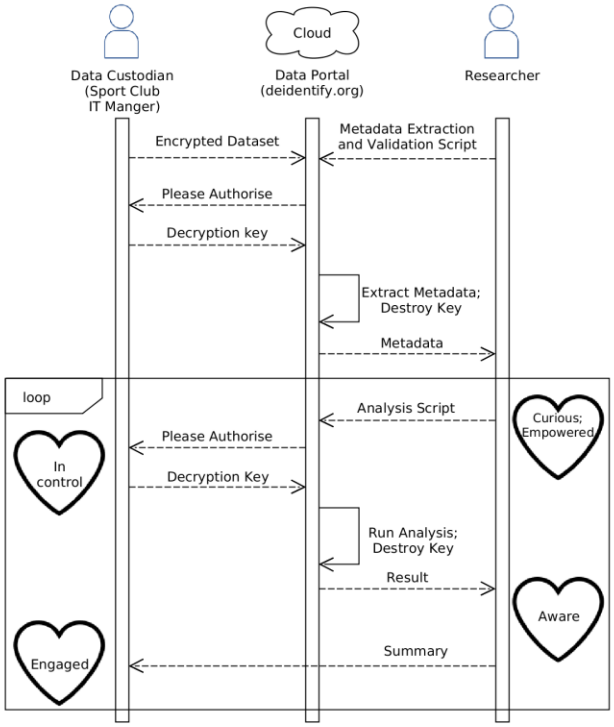

Figure 2: De-identification portal Interaction Model

4 IMPLEMENTATION

We implemented a proof of concept de-identification portal based on the above design. A cloud virtual machine was used to host the data portal, an AES 256-encrypted zip file to protect the dataset, Python script files for the researcher to express the proposed analysis, an HTML web interface for the data custodian to authorise a script by providing the decryption key, and a background process to run the analysis in the cloud and store the result.

5 HEURISTIC USABILITY EVALUATION

In Table 1 we perform a heuristic evaluation of our system according to the usability characteristics defined in the ISO software quality framework [12]. In contrast to prescriptive usability heuristic checklists such as those presented by Nielsen [13], our focus is on evaluating usability concerns stemming from the software architecture [14] rather than on minor usability issues that are implementation specific.

6 CASE STUDY

In this section we share our preliminary experience using the system to obtain de-identified player position tracking data from a sport club for team strategy analysis.

To deal with the unique privacy issues associated with human trajectory data, we selected a custom de-identification operation that combined downsampling the position data to 1 Hz with a randomly sorted point-cloud representation to increase uncertainty of player identities whenever two player paths crossed each other [15]. As our custom de-identification operation was too involved for the sport club to perform themselves, we asked the sport club to upload encrypted data to the de-identification portal described in this paper.

Implementing the analysis script to extract and de-identify data proved to be challenging without having a way to peek at the structure of the underlying raw data it was operating on. While theoretically this could be addressed through metadata or sample data, in this situation metadata was not available and the format of the sample data differed from the actual data. Thus, multiple iterations were necessary to infer the data structure and to address parsing related issues.

While ultimately successful, the overall process took one month as each iteration had to wait for manual authorisation by the sport club who acted as the data custodian.

7 CONCLUSIONS

Our tool and method are simpler for the data custodian at some additional burden to the researcher when compared against using a spreadsheet tool such as Microsoft Excel, the most commonly used tool for this operation currently. Specifically, our approach calls for the researcher to be able to express de-identification via an automated script. We argue this is a superior approach as the researcher would typically have additional skills and training to handle data compared to the data custodian.

While our case study examined sports club data, the approach is, in principle, generalizable to other domains in which the data custodian lacks the technical resources to de-identify the data themselves. Future work is needed to validate our emotional goal model through interviews with stakeholders, and to run an empirical trial to quantify the extent to which the system satisfies the emotional goals of the data custodian and researcher.

Future implementations could benefit from functionality to automatically reverse-engineer the structure and semantics of a dataset without revealing individuals in the dataset; this would reduce the number of iterations required for the researcher to understand the dataset and arrive at the final analysis, thus reducing risk of feelings of irritation and frustration from the data custodian and researcher.

Table 1: Heuristic Evaluation against ISO Usability Characteristics

ISO Usability Characteristic	ISO Definition	Spreadsheet	Ours
Appropriateness recognizability	Degree to which users can recognize whether a product or system is appropriate for their needs.	Spreadsheet editors are an obvious choice for data custodian to use to remove/substitute participant identifier columns, but data custodian may not be aware of need to also remove quasi-identifiers.	Researcher determines appropriate de-identification methods and sends link and instructions to data custodian.
Learnability	Degree to which a product or system can be used by specified users to achieve specified goals of learning to use the product or system with effectiveness, efficiency, freedom from risk and satisfaction in a specified context of use.	Spreadsheets provide a familiar and intuitive interface. However, without proper training, there is a risk of data errors due to incorrect formulas that refer to the wrong cells.	Researcher must have sufficient training to express de-identification operations.
Operability	Degree to which a product or system has attributes that make it easy to operate and control.	Spreadsheets provide an intuitive interface. However, may be slow and repetitive if need to manually apply the same operation to many worksheets.	The data custodian only needs to provide encryption/decryption password.
User error protection	Degree to which a system protects users against making errors.	Spreadsheet software has no intrinsic functionality for recognising identifiable data. Responsibility falls on data custodian.	Researcher can test their code on a sample data set. Data custodian's role is reduced to choosing an appropriate encryption password.
User interface aesthetics	Degree to which a user interface enables pleasing and satisfying interaction for the user.	Spreadsheet provides a familiar and intuitive interface.	Interface can be themed.
Accessibility	Degree to which a product or system can be used by people with the widest range of characteristics and capabilities to achieve a specified goal in a specified context of use.	Spreadsheets in default mode present issues for users with low vision.	Interface conforms to W3C Web Content Accessibility Guidelines.

REFERENCES

[1] National Health and Medical Research Council (NHMRC), and Australian Research Council (ARC), *National Statement on Ethical Conduct National Statement on Ethical Conduct in Human Research*, 2007 (Updated May 2015)

[2] L. Sweeney, "k-anonymity: a model for protecting privacy," *Int. J. Uncertainty, Fuzziness Knowledge-Based Syst.*, vol. 10, no. 05, pp. 557–570, 2002.

[3] K. LeFevre, D. J. D. J. DeWitt, and R. Ramakrishnan, "Incognito: efficient full-domain K-anonymity," *SIGMOD '05 Proc. 2005 ACM SIGMOD Int. Conf. Manag. data*, pp. 49–60, 2005.

[4] J. Brickell and V. Shmatikov, "The cost of privacy: destruction of data-mining utility in anonymized data publishing," *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discov. data Min.*, pp. 70–78, 2008.

[5] D. A. Norman, *Emotional design: Why we love (or hate) everyday things*. Basic Books, 2004.

[6] I. Ramos, D. M. Berry, and J. Á. Carvalho, "Requirements engineering for organizational transformation," *Inf. Softw. Technol.*, vol. 47, no. 7, pp. 479–495, 2005.

[7] A. Mendoza, J. Carroll, L. Stern, and S. Linda, "Software Appropriation Over Time : From Adoption to Stabilization and Beyond," *Australas. J. Inf. Syst.*, vol. 16, no. 2, pp. 5–23, 2010.

[8] A. Mendoza, T. Miller, S. Pedell, and L. Sterling, "The role of users' emotions and associated quality goals on appropriation of systems: Two case studies," *Proc. 24th Australas. Conf. Inf. Syst.*, 2013.

[9] M. K. Curumsing, N. Fernando, M. Abdelrazek, R. Vasa, K. Mouzakis, and J. Grundy, "Understanding the Impact of Emotions on Software: A Case Study in Requirements Gathering and Evaluation," *J. Syst. Softw.*, 2018.

[10] M. K. Curumsing, *Emotion-Oriented Requirements Engineering*, Swinburne University of Technology, 2017.

[11] P. Ohm, "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization," *UCLA Law Rev.*, vol. 57, no. 6, pp. 1701–1777, 2010.

[12] ISO/IEC 25010:2011, *Systems and software engineering—Systems and software Quality Requirements and Evaluation (SQuaRE)— System and software quality models*, 2011.

[13] J. Nielsen and R. Molich, "Heuristic Evaluation of user interfaces," *CHI '90 Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, April, pp. 249–256, 1990.

[14] E. Folmer and J. Bosch, "Architecting for usability: A survey," *J. Syst. Softw.*, vol. 70, no. 1–2, pp. 61–78, 2004.

[15] J. Ding, C.-C. Ni, and J. Gao, "Fighting Statistical Re-Identification in Human Trajectory Publication," in *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - SIGSPATIAL'17*, 2017