Minimalist approaches to enforce privacy by design in surveys

Enka Blanchard

Laboratoire d'Automatique, Mécanique et Informatique Industrielles et Humaines, UMR CNRS 8201, UPHF, CNRS,
Centre Internet et Société, UPR 2000, Centre National de la Recherche Scientifique

https://koliaza.com

Joint work with **L. Gabasova**

International Conference on Human-Computer Interaction RoCHI 2023, Cluj, 21-09-2023

A French case-study

Asked by a French university to organise a study on the quality of working environment:

- $\sim$ 120 invited participants (permanent faculty)
- $\sim$ 50 questions
- Some sensitive questions (medical/harassment)
- Strict legal (GDPR) and security rules

Setting goals: the research/practice gap

Similar studies in other French universities:

- Questionnaires written in an *ad hoc* fashion
- Commercial SaaS with no security guarantees
- Questions against the terms of use
- Relatively low response rates (∼20%)

Minimalism as a framework

Implementing privacy-by-design through minimalism:

- Separate checking the user's authorisation from their answers
- Avoid user-centred answer sheets, create a semi-independent database per question
- Compute only preregistered correlations, fully on the client-side
- Make the full list of questions and correlations public in advance

This reduces the organisers' powers but:

- Increases transparency, boosting participation and trust
- Limits the work done afterwards as preregistration makes it easier to automatise
- Pushes the organisers to ask whether they want to publish "everything is fine" or if they want to create a space to report real issues

Nominative information

Nominative information is generally unavoidable:

- Free-text answers
- Logins to access the survey
- Feedback on the survey

To handle:

- Make the authentication fungible (e.g., passphrases that can be exchanged)
- Store the authentication elsewhere
- Prevent any correlation with general free-text answers

Enforcing anonymity

To prevent deanonymisation, we get rid of answer sheets, hence:

- Store each question's answers as a single column
- Compute each correlation independently on the user side, have a column for each on the server
- Randomise the column's order after each insertion
- Prevent access to the database while the survey is active

Optionally

- At the end, automatically remove deanonymising questions
- Twin answers to avoid revealing a question's answers through its eventual removal

Self-identification questions

Problems with self-identification (e.g., for gender):

- Using inclusive language can be illegal and create political tension
- Not using it can also be illegal and create other problems
- It can lead to de-anonymisation

Potential solution:

- Make an open text field
- Parse answers according to a few categories for the correlations
- Keep both the full list and the categorised correlations

Open text fields

Drawbacks:

- Adds noise (depending on parser quality)
- More effort leads to higher non-response rate / lack of understanding
- Allows trolling

Advantages:

- Inclusive by design without making it visible
- Avoids the issues with having an “Other” option
- Compatible with evolving regulations
- Rare answers’ anonymity is protected by noise

Limits and perspectives

Limits:

- The system does not allow data modification and deletion
- Self-identification adds noise
- Self-imposed limits (for organisers) are hard to accept.

Open questions and future work:

- Can this be done without trusting the server?
- Could homomorphic encryption offer a practical alternative?
- How to best implement data modification?
- How does preregistration affect the responses?
- Could cookies be used to avoid user cost if the study must be restarted? Would it be worth the cost (including compliance with regulations such as GDPR)?