Project title

Automatic Detection of Identifiers in Open Data (ADIODA)

Lead applicant

Prof. dr. J.M. Jelte Wicherts

Affiliation

Tilburg University, Department of Methodology and Statistics

Team members

Chris Hartgerink, PhD - Liberate Science GmbH
Background: Meta-researcher specialized in detecting data fabrication and fostering responsible conduct of research, lead software engineer at Liberate Science, author of multiple R packages. Open scientist since 2012.
Roles: Develop and maintain the software produced during this project.

Richard Klein, Postdoc - Tilburg University
Background: Extensive experience in leading large, collaborative meta-science projects in psychology (Many Labs1, 2, 4), and co-project leader of the prior project assessing data privacy risks in open datasets (Wicherts et al., inprep).
Roles: coordinating data handling, supervising student assistants, studying the tool's performance, writingpublications

Jelte Wicherts, Professor - Tilburg University
Background: Extensive experience in leading meta-research projects on data management, and familiarity withexisting guidelines and ethical and legal frameworks regarding privacy. Co-led the prior project assessing data privacy risks in open datasets (Wicherts et al., in prep)
Roles: Overall project supervision, privacy literature review, writing publications, collaborations, dissemination

Public summary

Privacy breaches pose a major risk in the dissemination of rich datasets in the medical, social, and behavioural sciences, particularly when the data involve sensitive information. Here, we develop and validate an open tool called *Automatic Detection of Identifiers in Open Data (ADIODA)* allowing researchers in these fields to proactivelyand readily identify information in datasets that could (inadvertently) be used to (re-)identify individuals. ADIODA will be an open tool that can be easily implemented in research workflows, data audits, and editorial procedures to help protect the privacy of participants whose information is used in openly shared datasets.

Vision for the project

Researchers in the medical, social, and behavioural sciences increasingly share the data behind their scientific claims, with many benefits for society and science in terms of verification and re-use. Yet, when datasets contain sensitive information from human subjects, researchers need to be acutely aware of privacy risks of sharing data.

In a recent meta-research project funded by the European Research Council (Wicherts et al., in prep.), we assessed ~2,000 datasets published alongside articles in three popular psychology journals, where we observeddirect privacy risks (e.g., names, birth-dates, IP addresses) in about 5% of those datasets. Of those, over half contained sensitive data according to Europe Union's General Data Protection Regulation (GDPR). These ethicaland legal violations of privacy pose risks to participants, researchers, and universities.

The many benefits of open data mean that the solution to these risks can't be to reverse course and resume the outdated norm of closed data. With this grant, we hope to equip researchers with better safeguards to prevent privacy violations before they happen. Specifically, based on insights and data from our recent meta-research project, we aim to create and validate a tool called Automatic Detection of Identifiers in Open Data (ADIODA) thatallows researchers to readily identify possibly identifying information in datasets before they are shared. This open-source tool can be easily integrated into workflows to check whether the data researchers are about to disseminate contains common types of identifiable information according to the GDPR (EU) or the Health Insurance Portability and Accountability Act (USA). The tool is not meant to replace existing (rather bureaucratic) legal and organizational efforts dealing with privacy, but rather to offer an extra check that can be readily implemented to proactively lower the risks of privacy breaches in open data.

ADIODA is an important tool for creating responsible open data. The target audience includes all researchersstudying human subjects with quantitative data, including those who are still hesitant to share out of privacyconcerns.

The intended outputs are

- dictionary of empirically evaluated regular expressions to detect common privacy risks
- an implementation of these regular expressions in an R package
- an implementation in an easy-to-use desktop application for researchers to install on their device (sopotentially sensitive data does not need to be transmitted over the Internet).

We will implement a change management procedure to make researchers and editors of journals that include open data aware of this tool, such that we can work on bringing about social change. To this end, we will organizea virtual event to bring together researchers to discuss responsible data sharing.

This grant advances open science by working towards responsible data sharing, as a form of responsible researchconduct.

Open science track record

Jelte Wicherts has been an active proponent and vocal promotor of open science for over 15 years. He gave over 60 presentations on the benefits of open science across the globe, published over a dozen articles on data sharing,founded the Journal of Open Psychology Data, and sat in numerous (inter)national open science committees. At a policy level, he helped write the data management policy for the Dutch social science faculties (DSW) and chairs the science committee at his School.

He was organiser, invited participant, invited speaker of open science meetings all across the world, and obtained major grants (Vidi, NWO replication, ERC Consolidator) for his research on improving methodological rigor and transparency in (psychological) science. He helped develop the well-known tool Statcheck and developed the peerreview transparency tool underlying the Quality Open Access Market. His research on data sharing, replication, and errors & biases in statistics has had major impact (>11k citations). He obtained many international recognitions for his work, including being named a Fellow of the Association for Psychological Science, an early career award from the APA, becoming a member of John Ioannidis' Meta-Research Innovation Center at Stanford,and sitting in editorial boards of major journals.

Decision

Funded

Assessment report of the proposal

Motivation selection committee

Criterion 1: Quality of the project proposal (50%)

The committee judges that the aims and vision of the proposal fit very well with the aim of the Open Science Fund. The project's aim is to develop a tool to identify personal information in a dataset. The committee sees a lot of merit in the project because identifiable data is a significant problem that affects how much the public trusts the research community in terms of their privacy and their subsequent willingness to share data in the future. Tools are needed to ensure that inadvertent sharing of protected data does not happen, and therefore the committee concludes that the outputs of the project will be of great added value to implementing Open Science in practice.

Criterion 2: Feasibility of the project plan (40%)

The work plan is clearly described and appropriate for achieving the intended results. The committee especially values that good work has taken place previously, meaning that the project can move quickly and thus enhancing the feasibility of the project. It is unclear to the committee, however, how the tool will handle different types of datasets. The budget is adequate, but contains a discrepancy in the cost calculation of one of the budget posts (100 hours of work has been costed for instead of 150).

Criterion 3: Open Science track record of the main applicant (10%)
The main applicant has an excellent and notable Open Science track record and is active in several Open Science projects.